

Docket No. 0557-4645-2/art



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

IN RE APPLICATION OF: Tetsuro NAGATSUKA, et al.

GAU: 2786

SERIAL NO: 09/288,856

EXAMINER:

FILED: April 9, 1999

FOR: DOCUMENT CLASSIFICATION SYSTEM AND METHOD FOR CLASSIFYING A DOCUMENT ACCORDING TO CONTENTS OF THE DOCUMENT

REQUEST FOR PRIORITY

RECEIVED

JUL 16 1999

Group 2700

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231

SIR:

- Full benefit of the filing date of U.S. Application Serial Number [US App No], filed [US App Dt], is claimed pursuant to the provisions of 35 U.S.C. §120.
- Full benefit of the filing date of U.S. Provisional Application Serial Number , filed , is claimed pursuant to the provisions of 35 U.S.C. §119(e).
- Applicants claim any right to priority from any earlier filed applications to which they may be entitled pursuant to the provisions of 35 U.S.C. §119, as noted below.

In the matter of the above-identified application for patent, notice is hereby given that the applicants claim as priority:

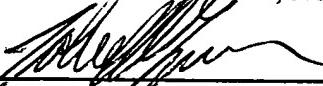
COUNTRY	APPLICATION NUMBER	MONTH/DAY/YEAR
JAPAN	10-114414	April 10, 1998
JAPAN	10-115907	April 13, 1998

Certified copies of the corresponding Convention Application(s)

- are submitted herewith
- will be submitted prior to payment of the Final Fee
- were filed in prior application Serial No. filed
- were submitted to the International Bureau in PCT Application Number .
Receipt of the certified copies by the International Bureau in a timely manner under PCT Rule 17.1(a) has been acknowledged as evidenced by the attached PCT/IB/304.
- (A) Application Serial No.(s) were filed in prior application Serial No. filed ; and
- (B) Application Serial No.(s)
 - are submitted herewith
 - will be submitted prior to payment of the Final Fee

Respectfully Submitted,

OBLON, SPIVAK, McCLELLAND,
MAIER & NEUSTADT, P.C.


Marvin J. Spivak
Registration No. 24,913
Robert F. Gnuse
Registration Number 27,905



日本国特許
PATENT OFFICE
JAPANESE GOVERNMENT

1999.4.23

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
る事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
in this Office.

出願年月日
Date of Application:

1998年 4月13日

RECEIVED

JUL 16 1999

願番号
Application Number:

平成10年特許願第115907号

Group 2700

願人
Applicant(s):

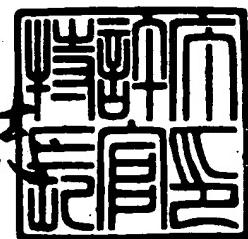
株式会社リコー

CERTIFIED COPY OF
PRIORITY DOCUMENT

1999年 4月23日

特許庁長官
Commissioner,
Patent Office

伴佐山建志



【書類名】 特許願
【整理番号】 9801994
【提出日】 平成10年 4月13日
【あて先】 特許庁長官殿
【国際特許分類】 G06F 17/27
G06F 17/30
【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体
【請求項の数】 33
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 剣持 栄治
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 宮地 達生
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 嶋田 敦夫
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 武谷 一寿
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 中島 明子
【発明者】
【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 長束 哲郎

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 山崎 真湖人

【発明者】

【住所又は居所】 東京都大田区中馬込1丁目3番6号 株式会社リコー内
【氏名】 藤田 克彦

【特許出願人】

【識別番号】 000006747
【氏名又は名称】 株式会社リコー
【代表者】 桜井 正光

【提出物件の目録】

【物件名】 明細書 1
【物件名】 図面 1
【物件名】 要約書 1

【書類名】 明細書

【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項1】 文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析し解析情報を得る解析手段と

前記解析手段により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出手段と、

前記変換関数算出手段により算出された表現空間変換関数をもちいて前記ベクトル生成手段により生成された文書特徴ベクトルを変換するベクトル変換手段と

前記ベクトル変換手段により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類手段と、

前記分類手段により分類された文書分類の結果を記憶する分類結果記憶手段と

を備えたことを特徴とする文書分類装置。

【請求項2】 前記ベクトル生成手段により生成された文書特徴ベクトル間の内積を算出する内積算出手段を備え、

前記変換関数算出手段は、前記内積算出手段により算出された内積をもちいて表現空間変換関数を算出することを特徴とする請求項1に記載の文書分類装置。

【請求項3】 前記入力手段により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する文書間類似情報設定手段を備え、

前記変換関数算出手段は、前記内積算出手段により算出された内積および前記文書間類似情報設定手段により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする請求項2に記載の文書分類装置。

【請求項4】 さらに、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶するベクトル記憶手段と、

前記変換関数算出手段により算出された表現空間変換関数を記憶する変換関数記憶手段と、

を備えたことを特徴とする請求項1～3のいずれか一つに記載された文書分類装置。

【請求項5】 さらに、前記ベクトル変換手段により文書特徴ベクトルを変更する前に、前記解析手段により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正するベクトル修正手段を備えたことを特徴とする請求項1～4のいずれか一つに記載された文書分類装置。

【請求項6】 前記ベクトル修正手段において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により前記ベクトル変換手段において前記文書特徴ベクトルが適切に変換できるように、前記変換関数算出手段により算出された表現空間変換関数を修正する変換関数修正手段を備えたことを特徴とする請求項5に記載の文書分類装置。

【請求項7】 さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする変換関数修正指示手段と、

前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する変換関数修正手段と、

を備えたことを特徴とする請求項1～5のいずれか一つに記載の文書分類装置。

【請求項8】 前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項7に記載の文書分類装置。

【請求項9】 前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項7に記載の文書分類装置。

【請求項10】 前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記解析手段により得られた解析情報をもついて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項7に記載の文書分類装置。

【請求項11】 前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記分類結果記憶手段により記憶された分類結果をもついて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項7に記載の文書分類装置。

【請求項12】 初期クラスタ重心を指定する初期重心指定手段と、前記初期重心指定手段により指定された初期クラスタ重心を登録する初期重心登録手段とを備え、

前記分類手段は、前記初期重心登録手段により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする請求項1～11のいずれか一つに記載の文書分類装置。

【請求項13】 前記初期重心指定手段により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする請求項12に記載の文書分類装置。

【請求項14】 前記初期重心指定手段により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする請求項12に記載の文書分類装置。

【請求項15】 前記初期重心指定手段により指定される初期クラスタ重心として前記解析手段により得られた解析情報を指定することを特徴とする請求項12に記載の文書分類装置。

【請求項16】 前記初期重心指定手段により指定される初期クラスタ重心として前記分類結果記憶手段により記憶された分類結果を指定することを特徴とする請求項12に記載の文書分類装置。

【請求項17】 文書データを入力する第1工程と、前記第1工程により入力された文書データを解析し解析情報を得る第2工程と

前記第2工程により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成する第3工程と、

前記第3工程により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する第4工程と、

前記第4工程により算出された表現空間変換関数をもちいて前記第3工程により生成された文書特徴ベクトルを変換する第5工程と、

前記第5工程により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する第6工程と、

前記第6工程分類手段により分類された文書分類の結果を記憶する第7工程と

を含んだことを特徴とする文書分類方法。

【請求項18】 前記第3工程により生成された文書特徴ベクトル間の内積を算出する第8工程を含み、

前記第4工程は、前記第8工程により算出された内積をもちいて表現空間変換関数を算出することを特徴とする請求項17記載の文書分類方法。

【請求項19】 前記第1工程により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する第9工程を含み、

前記第4工程は、前記第8工程により算出された内積および前記第9工程により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする請求項18に記載の文書分類方法。

【請求項20】 さらに、前記第3工程により生成された文書特徴ベクトルを記憶する第10工程と、

前記第4工程により算出された表現空間変換関数を記憶する第11工程と、
を含んだことを特徴とする請求項17～19のいずれか一つに記載された文書分類方法。

【請求項21】 さらに、前記第5工程により文書特徴ベクトルを変更する前に、前記第2工程により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正する第12工程を含んだ

ことを特徴とする請求項17~20のいずれか一つに記載された文書分類方法。

【請求項22】 前記第12工程において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により第5工程において前記文書特徴ベクトルが適切に変換できるように、前記第4工程により算出された表現空間変換関数を修正する第13工程を含んだことを特徴とする請求項21に記載の文書分類方法。

【請求項23】 さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする第14工程と、

前記第14工程により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する第15工程と、
を含んだことを特徴とする請求項17~21のいずれか一つに記載の文書分類方法。

【請求項24】 前記第15工程により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもつて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項23に記載の文書分類方法。

【請求項25】 前記第15工程により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもつて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項23に記載の文書分類方法。

【請求項26】 前記第15工程により指示された特徴次元の操作に関する指示内容が、前記第2工程により得られた解析情報をもつて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項23に記載の文書分類方法。

【請求項27】 前記第15工程により指示された特徴次元の操作に関する指示内容が、前記第7工程により記憶された分類結果をもつて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする請求項23に記載の文書分類方法。

【請求項28】 初期クラスタ重心を指定する第16工程と、

前記第16工程により指定された初期クラスタ重心を登録する第17工程とを含み、

前記第6工程は、前記第17工程により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする請求項17~27のいずれか一つに記載の文書分類方法。

【請求項29】 前記第16工程により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする請求項28に記載の文書分類方法。

【請求項30】 前記第16工程により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする請求項28に記載の文書分類方法。

【請求項31】 前記第16工程により指定される初期クラスタ重心として前記第2工程により得られた解析情報を指定することを特徴とする請求項28に記載の文書分類方法。

【請求項32】 前記第16工程により指定される初期クラスタ重心として前記第7工程により記憶された分類結果を指定することを特徴とする請求項28に記載の文書分類方法。

【請求項33】 前記請求項17~32のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は、文書間の類似性に基づいて文書を分類する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【0002】

【従来の技術】

近年インターネット等の普及により大量の文書情報へのアクセスが可能になり、収集した大量の文書情報を意味のあるカテゴリーに分類し、文書群の構造を把握するなどの知的作業がおこなわれるようになっていきている。大量の文書情報

を操作者が手動で分類する場合、人的／時間的コストが膨大なものになり、また分類をおこなう際にもちいる知識を分類をする操作者のみが有することになるため、分類をおこなう担当の操作者が変わると分類基準も変わってしまうことになる。

【0003】

したがって、文書群をいかに人間が分類をおこなうような分類基準によって自動的に分類することができるかが重要な課題となる。すなわち、意味的に類似している文書は同一のカテゴリーに分類され、また、分類をする工程において生成される各分類カテゴリーは操作者が文類実行前に意図しているような分類カテゴリーとなるように構成された文書分類装置の出現が望まれている。

【0004】

文書の自動分類装置の従来技術としては、たとえば特開平7-36897号公報に記載されているように、文書を単語を特徴とする文書ベクトルとみなし、クラスタリング手法をもちいてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【0005】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、分類対象文書に含まれる単語を特徴量とする文書特徴ベクトルをもちいて、その文書特徴ベクトルに対しクラスタリング手法を適用して分類をおこなうため、単語の多義性／同義性により文書の意味的な関連性を反映した分類結果を得ることが困難となるという問題があった。

【0006】

この単語の多義性／同義性の問題を解決するものとしては、米国特許第4839853号公報に記載されているように、文書間の内積行列に特異値分解を適用するものがある。すなわち、文書間の単語の共起性をもとに生成される潜在的意味空間といわれる空間へ、文書と単語を射影することにより意味的な関連性を反映した文書検索をおこなうものである。

【0007】

また、「Projections for Efficient Document Clustering (著者名: Hinrich Schutze and Craig Silverstein, 学会名: ACM, 論文名: Proceedings of SIGIR, ページ: 74-81, 発行年: 1997)」においては、上記潜在的意味空間において文書分類を実施しているものがある。さらに、「Representating Documents Using an Explicit Model of Their Similarities (著者名: Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew, 論文名: Journal of the American Society for Information Science, 学会名: the American Society for Information Science, ページ: 254-271, Vol. 46 No. 4, 発行年: 1995)」においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有する他の文書への参照情報から生成される共参照情報などを付加した行列をもちいて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものがある。

【0008】

これらの従来技術の手法で生成される射影空間の各次元は複数の単語が意味的に結合した概念的なものであるが、どの特徴次元を使って文書分類あるいは文書検索をおこなうかは、特異値分解を適用する際に算出される特異値の大きさのみを基準として決定される。このため、分類実行時にもちいられる特徴次元の選択においては、操作者の意図が反映されることは困難であり、このため分類結果が操作者の意図する結果と異なってしまうという問題点があった。

【0009】

また、従来の他の文書分類方法では、文書の意味的な関連性を反映した文書分類をおこなうために、文書を変換するための表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書の文書分類をおこなう部分とを連続的に処理しているが、表現空間変換関数を算出する部分は非常に

計算時間を費やす処理であるため、結果として一回の文書分類に要する時間も膨大なものになるという問題点があった。

【0010】

この発明は、上述した従来例による問題点を解消するため、操作者の意図を反映する文書分類を短時間で効率良く繰り返しをおこなうことができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【0011】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項1の発明に係る文書分類装置は、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析し解析情報を得る解析手段と、前記解析手段により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出手段と、前記変換関数算出手段により算出された表現空間変換関数をもつて前記ベクトル生成手段により生成された文書特徴ベクトルを変換するベクトル変換手段と、前記ベクトル変換手段により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類手段と、前記分類手段により分類された文書分類の結果を記憶する分類結果記憶手段と、を備えたことを特徴とする。

【0012】

この請求項1の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能である。

【0013】

また、請求項2に係る文書分類装置は、請求項1の発明において、前記ベクトル生成手段により生成された文書特徴ベクトル間の内積を算出する内積算出手段

を備え、前記変換関数算出手段が、前記内積算出手段により算出された内積をもちいて表現空間変換関数を算出することを特徴とする。

【0014】

この請求項2の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【0015】

また、請求項3に係る文書分類装置は、請求項2の発明において、前記入力手段により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する文書間類似情報設定手段を備え、前記変換関数算出手段が、前記内積算出手段により算出された内積および前記文書間類似情報設定手段により設定された文書間類似情報をもちいて表現空間変換関数を算出することを特徴とする。

【0016】

この請求項3の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【0017】

また、請求項4に係る文書分類装置は、請求項1～3の発明において、さらに、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶するベクトル記憶手段と、前記変換関数算出手段により算出された表現空間変換関数を記憶する変換関数記憶手段と、を備えたことを特徴とする。

【0018】

この請求項4の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数を

もちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能である。

【0019】

また、請求項5に係る文書分類装置は、請求項1～4のいずれか一つの発明において、さらに、前記ベクトル変換手段により文書特徴ベクトルを変更する前に、前記解析手段により抽出される単語が有する特性により構成される規則をもついて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正するベクトル修正手段を備えたことを特徴とする。

【0020】

この請求項5の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能である。

【0021】

また、請求項6に係る文書分類装置は、請求項5の発明において、前記ベクトル修正手段において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により前記ベクトル変換手段において前記文書特徴ベクトルが適切に変換できるように、前記変換関数算出手段により算出された表現空間変換関数を修正する変換関数修正手段を備えたことを特徴とする。

【0022】

この請求項6の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもついて変換された文書をもついて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能である。

【0023】

また、請求項7に係る文書分類装置は、請求項1～5のいずれか一つの発明において、さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする変

換関数修正指示手段と、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する変換関数修正手段と、を備えたことを特徴とする。

【0024】

この請求項7の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0025】

また、請求項8に係る文書分類装置は、請求項7の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0026】

この請求項8の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0027】

また、請求項9に係る文書分類装置は、請求項7の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0028】

この請求項9の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0029】

また、請求項10に係る文書分類装置は、請求項7の発明において、前記変換

関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記解析手段により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0030】

この請求項10の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0031】

また、請求項11に係る文書分類装置は、請求項7の発明において、前記変換関数修正指示手段により指示された特徴次元の操作に関する指示内容が、前記分類結果記憶手段により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0032】

この請求項11の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0033】

また、請求項12に係る文書分類装置は、請求項1～11のいずれか一つの発明において、初期クラスタ重心を指定する初期重心指定手段と、前記初期重心指定手段により指定された初期クラスタ重心を登録する初期重心登録手段とを備え、前記分類手段は、前記初期重心登録手段により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする。

【0034】

この請求項12の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0035】

また、請求項13に係る文書分類装置は、請求項12の発明において、前記初期重心指定手段により指定される初期クラスタ重心として任意の文書ベクトルデータを指定することを特徴とする。

【0036】

この請求項13の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0037】

また、請求項14に係る文書分類装置は、請求項12の発明において、前記初期重心指定手段により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする。

【0038】

この請求項14の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0039】

また、請求項15に係る文書分類装置は、請求項12の発明において、前記初期重心指定手段により指定される初期クラスタ重心として前記解析手段により得られた解析情報を指定することを特徴とする。

【0040】

この請求項15の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0041】

また、請求項16に係る文書分類装置は、請求項12の発明において、前記初

期重心指定手段により指定される初期クラスタ重心として前記分類結果記憶手段により記憶された分類結果を指定することを特徴とする。

【0042】

この請求項16の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0043】

また、請求項17に係る文書分類方法は、文書データを入力する第1工程と、前記第1工程により入力された文書データを解析し解析情報を得る第2工程と、前記第2工程により得られた解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成する第3工程と、前記第3工程により生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する第4工程と、前記第4工程により算出された表現空間変換関数をもちいて前記第3工程により生成された文書特徴ベクトルを変換する第5工程と、前記第5工程により変換された文書特徴ベクトル間の類似度に基づいて文書を分類する第6工程と、前記第6工程分類手段により分類された文書分類の結果を記憶する第7工程と、を含んだことを特徴とする。

【0044】

この請求項17の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能である。

【0045】

また、請求項18に係る文書分類方法は、請求項17の発明において、前記第3工程により生成された文書特徴ベクトル間の内積を算出する第8工程を含み、前記第4工程は、前記第8工程により算出された内積をもちいて表現空間変換関数を算出することを特徴とする。

【0046】

この請求項18の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【0047】

また、請求項19に係る文書分類方法は、請求項18の発明において、前記第1工程により入力された文書の作成者、作成日等の文書データの文書間類似情報を設定する第9工程を含み、前記第4工程は、前記第8工程により算出された内積および前記第9工程により設定された文書間類似情報をもついて表現空間変換関数を算出することを特徴とする。

【0048】

この請求項19の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能である。

【0049】

また、請求項20に係る文書分類方法は、請求項17～19のいずれか一つの発明において、さらに、前記第3工程により生成された文書特徴ベクトルを記憶する第10工程と、前記第4工程により算出された表現空間変換関数を記憶する第11工程と、を含んだことを特徴とする。

【0050】

この請求項20の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもついて変換された文書をもついて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能である。

【0051】

また、請求項21に係る文書分類方法は、請求項17～20のいずれか一つの発明において、さらに、前記第5工程により文書特徴ベクトルを変更する前に、前記第2工程により抽出される単語が有する特性により構成される規則をもちいて前記文書特徴ベクトルおよび／または文書特徴ベクトルを構成する特徴次元を操作することにより前記文書特徴ベクトルを修正する第12工程を含んだことを特徴とする。

【0052】

この請求項21の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能である。

【0053】

また、請求項22に係る文書分類方法は、請求項21の発明において、前記第12工程において文書特徴ベクトルを修正することにより特徴次元が変更された場合に、前記変更された特徴次元により第5工程において前記文書特徴ベクトルが適切に変換できるように、前記第4工程により算出された表現空間変換関数を修正する第13工程を含んだことを特徴とする。

【0054】

この請求項22の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能となる。

【0055】

また、請求項23に係る文書分類方法は、請求項17～21のいずれか一つの発明において、さらに、前記表現空間変換関数の特徴次元の操作に関する指示をする第14工程と、前記第14工程により指示された特徴次元の操作に関する指示内容に基づいて、前記表現空間変換関数を修正する第15工程と、を含んだことを特徴とする。

【0056】

この請求項23の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0057】

また、請求項24に係る文書分類方法は、請求項23の発明において、前記第15工程により指示された特徴次元の操作に関する指示内容が、任意の文書ベクトルデータをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0058】

この請求項24の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0059】

また、請求項25に係る文書分類方法は、請求項23の発明において、前記第15工程により指示された特徴次元の操作に関する指示内容が、文書特徴ベクトルをもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0060】

この請求項25の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0061】

また、請求項26に係る文書分類方法は、請求項23の発明において、前記第15工程により指示された特徴次元の操作に関する指示内容が、前記第2工程により得られた解析情報をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0062】

この請求項26の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0063】

また、請求項27に係る文書分類方法は、請求項23の発明において、前記第15工程により指示された特徴次元の操作に関する指示内容が、前記第7工程により記憶された分類結果をもちいて前記表現空間変換関数の特徴次元を操作するものであることを特徴とする。

【0064】

この請求項27の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能である。

【0065】

また、請求項28に係る文書分類方法は、請求項17～27のいずれか一つの発明において、初期クラスタ重心を指定する第16工程と、前記第16工程により指定された初期クラスタ重心を登録する第17工程とを含み、前記第6工程は、前記第17工程により登録された初期クラスタ重心にしたがって文書を分類することを特徴とする。

【0066】

この請求項28の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0067】

また、請求項29に係る文書分類方法は、請求項28の発明において、前記第16工程により指定される初期クラスタ重心として任意の文書ベクトルデータを

指定することを特徴とする。

【0068】

この請求項29の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0069】

また、請求項30に係る文書分類方法は、請求項28の発明において、前記第16工程により指定される初期クラスタ重心として文書特徴ベクトルを指定することを特徴とする。

【0070】

この請求項30の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0071】

また、請求項31に係る文書分類方法は、請求項28の発明において、前記第16工程により指定される初期クラスタ重心として前記第2工程により得られた解析情報を指定することを特徴とする。

【0072】

この請求項31の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0073】

また、請求項32に係る文書分類方法は、請求項28の発明において、前記第16工程により指定される初期クラスタ重心として前記第7工程により記憶された分類結果を指定することを特徴とする。

【0074】

この請求項32の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能である。

【0075】

また、請求項33の発明に係る記憶媒体は、請求項17~32に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項17~32の動作をコンピュータによって実現することが可能である。

【0076】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【0077】

(実施の形態1)

まず、この発明の実施の形態1による文書分類装置を構成する情報処理システム全体のハードウェア構成を説明する。図1は、実施の形態1による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【0078】

図1において、実施の形態1による文書分類装置を構成する情報処理システムは、サーバー/クライアント方式で構成されている。すなわち、サーバー101と複数のクライアント102がネットワーク103によって接続されている。クライアント102は、分類データの生成、サーバー101への指示、分類結果の表示などをおこなう。一方、クライアント102からの指示にしたがって、サーバー101は文書(テキスト)分類に関する処理を膨大な数値演算によりおこない、その処理の結果をクライアント102へ送る。

【0079】

より具体的には、サーバー101においては、テキスト分類処理(前処理、ク

ラスタリング処理) がおこなわれ、クライアント102においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー101における処理は、上述のように、「前処理」と「分類処理」の2つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー101は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。「前処理」および「分類処理」の詳細については後述する。

【0080】

また、サーバー101とクライアント102との間のデータのやりとりはファイル共有という方法をもちいる。すなわち、分類処理にもちいるファイルをサーバー101上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント102からはサーバー101の共有フォルダをネットワーク共有して利用することが可能である。

【0081】

つぎに、サーバー101およびクライアント102のハードウェア構成について説明する。図2は、実施の形態1による文書分類装置を構成する情報処理システムにおけるサーバー101をハードウェア的に示す説明図である。サーバー101は、たとえばワークステーション(WS)等がもちいられる。

【0082】

図2において、201はサーバー101全体を制御するCPUを、202はポートプログラム等を記憶したROMを、203はCPU201のワークエリアとして使用されるRAM203を、204は通信回線205を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス(I/F)を、206はデータを記憶するディスク装置を示している。200は上記各部を結合させるためのバスを示している。

【0083】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ208や、データを入力するためのキーボード209およびマウス210等が同様に接続されていてもよい。さらに、ディスク装置206には、クライアント102との

間のデータのやりとりをするための共有フォルダ207が設けられている。

【0084】

また、図3は、実施の形態1による文書分類装置を構成する情報処理システムにおけるクライアント102をハードウェア的に示す説明図である。クライアント102は、たとえばパーソナルコンピュータ(PC)等がもちいられる。

【0085】

図3において、301はシステム全体を制御するCPUを、302はブートプログラム等を記憶したROMを、303はCPU301のワークエリアとして使用されるRAMを、304はCPU301の制御にしたがってHD(ハードディスク)305に対するデータのリード/ライトを制御するHDD(ハードディスクドライブ)を、305はHDD304の制御で書き込まれたデータを記憶するHDを、306はCPU301の制御にしたがってFD(フロッピーディスク)307に対するデータのリード/ライトを制御するFDD(フロッピーディスクドライブ)を、307はFDD306の制御で書き込まれたデータを記憶する着脱自在のFDを、308はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【0086】

また、309は通信回線310を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス(I/F)を、311は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、312はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、313はOCR(Optical Character Reader)機能を備えた画像を光学的に読み取るスキャナを、314は分類結果を含むデータの内容等を印刷するプリンタを、315は上記各部を結合するためのバスをそれぞれ示している。また、HD305にはワープロソフトや表計算ソフト等のアプリケーションソフト316が記憶されている。

【0087】

つぎに、実施の形態1による文書分類装置の機能的構成について説明する。図

4～図6は、実施の形態1による文書分類装置の構成を機能的に示すブロック図である。図4において、文書分類装置は、入力部401と、解析部402と、ベクトル生成部403と、変換関数算出部404と、ベクトル変換部405と、分類部406と、分類結果記憶部407を含む構成である。

【0088】

さらに、入力部401と解析部402との間には、文書データ中の表記の揺れ等を吸収する図示しない第1フィルタ部を含めるようにしてもよい。また、解析部402とベクトル生成部403との間には、解析情報から不要な単語や品詞を除去する図示しない第2フィルタ部を含めるようにしてもよい。さらに、変換関数算出部404とベクトル変換部との間には、文書特徴ベクトルから分類時に不要な単語や品詞を除去する図示しない第3フィルタ部を含めるようにしてもよい。

【0089】

また、図5においては、さらに内積算出部421を含む構成となっている。また、図6においては、さらに文書間類似情報設定部431を含む構成となっている。

【0090】

入力部401は、文書データを入力するものであり、たとえば、キーボード209または311、スキャナ313、OCR機能を備えたスキャナ313、またはネットワーク103を経由して文書や文書群を得ることができるI/F204または309等である。また、入力部401は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。さらに、入力した文書データを記憶する図示しない文書データ記憶部を含んでいてもよい。

【0091】

ここで、文書とは、自然言語で記述された一つ以上の文の集まりであり、それが分類対象となる場合はこれを文書という。具体的には、公開特許公報や特定の

新聞記事も文書であり、また、請求項や特定の一文を取り出したものであっても、これを文書とみなすものである。

【0092】

解析部402は、入力部401により入力された文書データの単語を解析し解析情報を得る。具体的には、入力部401により入力された文書データそれぞれに対して、形態素解析等の自然言語解析をおこない、単語やその品詞などを抽出する。さらに、文書群で出現した単語に対し一意な単語IDを付与し、文書内および文書群に対する単語出現回数を計数するものである。

【0093】

また、ベクトル生成部403は、解析部403により得られた解析情報に基づいて文書データに対する文書特徴ベクトルを生成するものである。変換関数算出部404は、文書特徴ベクトル間の類似性を反映する空間に前記ベクトル生成部403により生成された文書特徴ベクトルを射影するための表現空間変換関数を算出するものである。ベクトル変換部405は、変換関数算出部404により算出された表現空間変換関数をもちいて文書特徴ベクトルを変換するものである。

【0094】

ベクトル生成部403、変換関数算出部404、ベクトル変換部405の各処理の詳細は後述する。

【0095】

分類部406は、ベクトル変換部405により変換された新たな文書特徴ベクトル間の類似度に基づいて文書を分類するものである。具体的には、生成された分類対象データに対して、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等の分類手法を適用することで、文書分類をおこなうことができる。分類部406においては、ベクトルデータが適用できる分類手法であれば、その手法は問わない。

【0096】

さらに、分類結果記憶部407は、分類部406により分類された結果を適切な形式で記憶する記憶部である。たとえば、ディスク装置306またはハードディスク316の所定の領域のほか、RAM203または303、その他データを

記憶可能なところであればいずれでもよい。

【0097】

内積算出部501は、ベクトル生成部402手段により生成された文書特徴ベクトル間の内積を算出する算出部である。内積算出部501の処理の内容は後述する。

【0098】

また、文書間類似情報設定部601は、入力部401により入力された文書の作者、作成日等の文書データの文書間類似情報を設定する設定部である。文書間類似情報には、文書内での単語の出現順序や、文書の作成日、修正日、作者、修正者、参照文書、引用文書などの文書間での一致情報を含む。操作者は、これらの文書間類似情報の中から所望の情報を指定し任意に設定することができる。

【0099】

入力部401、解析部402、ベクトル生成部403、変換関数算出部404、ベクトル変換部405、分類部406、内積算出部420、文書間類似情報設定部430は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0100】

つぎに、ベクトル生成部403による文書特徴ベクトルの生成処理の内容について説明する。ベクトル生成部403は、解析部403により得られた解析情報に基づいて文書データに対する文書特徴ベクトルを生成するものである。ここで解析情報とは、たとえば、単語、単語ID、単語出現回数、品詞情報等の情報である。

【0101】

図7は、文書-単語行列データと文書特徴ベクトルの一例を示す説明図である。図7において、行成分701が単語IDであり、また、列成分702が文書IDである。行列要素として、文書IDが列番号であり、文書に含まれる単語ID

が行番号である単語の出現回数となるような文書－単語行列データを上記解析情報に基づいて生成する。この文書－単語行列の各列ベクトルが文書特徴ベクトルである。このようにして文書特徴ベクトルを生成する。

【0102】

また、この文書特徴ベクトルに対して、正規化等の処理を同時におこなうこともできる。この際、文書－単語行列データに付随する付加的な情報、たとえば、文書－単語行列データの行成分である単語IDとその単語との対応関係を記述した単語－単語ID対応マップデータや各単語において単語IDとその単語が有する品詞情報との対応関係を記述した単語ID－品詞対応マップデータなども同時に生成する。

【0103】

つぎに、変換関数算出部404による変換関数算出処理の内容について説明する。ベクトル生成部403における文書特徴ベクトルの生成は、通常、その文書内での単語の出現回数に基づいておこなわれる。この際、各単語はそれぞれ意味的に独立なものと仮定し、各々を直交するものとして扱われる。しかしながら、現実には単語は多義性や同義性を含むものであるため、上記のような仮定の妥当性は保証されておらず、各単語が各々直交するものと扱われることにより、分類の精度・妥当性にも影響を及ぼすものである。

【0104】

この影響を軽減するための手法として、この問題を多次元尺度問題とみなして、統計的手法をもちいることが考えられる。すなわち、変換関数算出部404において、各文書特徴ベクトルを文書特徴ベクトル間での特徴次元、すなわち単語の共起性が反映された空間へ変換するための表現空間変換関数を、ベクトル生成部403により生成された文書特徴ベクトルに基づいて算出する。なお、単語間の同義性の影響を軽減するための方法としてシソーラス等をもちいるようにしてもよい。

【0105】

本実施の形態においては表現空間変換関数の算出手法としては、前出の「*Representing Documents Using an Expl*

icit Model of Their Similarities」に述べられている表現空間変換関数の算出手法をもちいるが、そのほか、因子分析や数量化などの手法をもちいて算出するようにしてもよい。

【0106】

すなわち、内積算出部501により算出された文書特徴ベクトル間の内積に、文書間類似情報設定手段により設定された文書間類似情報を付加した文書間類似行列と文書特徴ベクトルに基づいて表現空間変換関数を算出する。そして、この表現空間変換関数をもちいることにより、文書間の意味的な類似性を強く反映した表現空間にて文書分類をおこなうことができる。また、上述のように、操作者が自由に文書間類似情報を付加的に選択することもできるため、操作者の意図を反映した文書分類をおこなうことができる。

【0107】

具体的には、文書数をd、単語数をtとし、大きさ $t \times d$ の文書－単語行列（文書特徴ベクトル）をX、大きさ $d \times d$ の文書間内積行列をS、大きさ $d \times d$ の付加的文書間類似情報行列を S_a とすると、表現空間変換関数Wは式1のようになる。

【0108】

$$W = M^T C X^+ \quad (\text{式1})$$

【0109】

なお、^Tは行列の転置を示す。

【0110】

ここで、行列へ特異値分解を適用する演算子を $s \circ d$ () とすると、式1において、行列C、M、 X^+ はつぎのような行列となる。

【0111】

$$X = s \circ d (X) = U L A^T \quad (\text{式2})$$

【0112】

$$S = X^T X \quad (\text{式3})$$

【0113】

$$S + S_a = s \circ d (S + S_a) = C^T C \quad (\text{式4})$$

【0114】

$$CAA^T = svd(CAA^T) = MZN^T \quad (\text{式 } 5)$$

【0115】

$$X^+ = A L^{-1} U^T \quad (\text{式 } 6)$$

【0116】

また、ベクトルの内積をもちいて表現空間変換関数を算出するには、上記付加的文書間類似行列 S_a を空行列とする。その場合、表現空間変換関数は式7によくになる。

【0117】

$$W = U^T \quad (\text{式 } 7)$$

【0118】

また、設定された文書間類似情報をもちいて表現空間変換関数を算出するには、 S_a を空行列以外の対称行列とする。

【0119】

さらに、本文書分類装置では、表現空間変換関数Wを大きさが $t \times t$ の単位行列とすることで変換関数生成部404をバイパスすることも可能である。

【0120】

さらにまた、ベクトル生成部403で生成される文書特徴ベクトルは、特徴次元数が文書群で出現する単語数であるため、通常非常に高次なものとなり、このまま分類等をおこなうと計算コストや記憶空間が膨大になる。このため、出現回数の極端に少ない単語や極端に多い単語を文書特徴ベクトルを構成する次元から除外することができるが、これにより分類精度や妥当性が低下する可能性がある。

【0121】

本発明でもちいる表現空間変換関数は各文書特徴ベクトル間の単語の共起性が考慮された空間への変換を実現するため、式1からも明らかのように表現空間変換関数により生成される表現空間は各特徴次元が複数の単語の一次結合として表現される。したがって、少ない特徴次元でも多くの単語の意味を扱うことができ、これにより分類等をおこなう際の計算コストや記憶空間を抑制することができ

る。

【0122】

つぎに、ベクトル変換部405による文書特徴ベクトルの変換処理について説明する。ベクトル変換部405では、変換関数生成部404で生成される表現空間変換関数をもちいて、文書特徴ベクトルを変換し、分類の対象となるデータを導く。加えて、各単語も前記表現空間変換関数をもちいて変換するが可能である。すなわち、表現空間変換関数として行列Wをもちいると、変換された文書特徴ベクトルをD_p とすると、式8のようになる。

【0123】

$$D_p = W X \quad (\text{式8})$$

【0124】

また、変換された単語の行列表現をT_p とすると、式9のようになる。

【0125】

$$T_p = W^T I = W \quad (\text{式9})$$

【0126】

なお、Iは単位行列を示す。

【0127】

つぎに、実施の形態1による文書分類装置の一連の処理の手順について説明する。図8は実施の形態1による文書分類装置の一連の処理の手順を示すフローチャートである。図8のフローチャートにおいて、まず、入力部401は文書データを入力する（ステップS810）。つぎに、解析部402はステップS810において入力された文書データを解析し解析情報を得る（ステップS820）。

【0128】

つぎに、ベクトル生成部403はステップS820において得解析情報に基づいて文書特徴ベクトルを生成する（ステップS830）。つぎに、変換関数算出部404はステップS830において生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間関数を算出する（ステップS840）。

【0129】

つぎに、ベクトル変換部405はステップS840において算出された表現空間関数をもちいてステップS830において生成された文書特徴ベクトルを変換する（ステップS850）。つぎに、分類部406はステップS850において変換された文書特徴ベクトルの間の類似度に基づいて文書を分類する（ステップS860）。その後、ステップS860によって分類された分類結果が記憶され（ステップS870）、すべての処理を終了する。

【0130】

また、図9は実施の形態1による文書分類装置の一連の処理の別の手順を示すフローチャートである。図9のフローチャートにおいて、図8の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【0131】

ステップS830につづいて、同ステップにおいて生成された文書特徴ベクトル間の内積を算出する（ステップS835）。つぎに、文書間類似情報をもちいるとの指示があったか否かを判断する（ステップS836）。

【0132】

ステップS836において、指示がなかった場合（ステップS836否定）は、ステップS835において算出された内積をもちいて表現空間変換関数の算出をする（ステップS840）。一方、ステップS836において、指示があった場合（ステップS836肯定）は、入力部401により入力された文書データの文書間類似情報を設定する（ステップS837）。その後、ステップS840へ移行し、ステップS835において算出された内積とステップS837において設定された文書類間情報をもちいて表現空間変換関数の算出をする。以下、図8と同様の処理をおこなう。

【0133】

以上説明したように、実施の形態1によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【0134】

(実施の形態2)

さて、上述した実施の形態1では、生成された文書特徴ベクトルと算出された表現空間変換関数の保存についてはなんら記載していなかったが、以下に説明する実施の形態2のように、さらにベクトル記憶部と、変換関数記憶部とを含む構成とするようにしてもよい。

【0135】

実施の形態2による文書分類装置の機能的構成について説明する。図10は、実施の形態2による文書分類装置の構成を機能的に示すブロック図である。図10において、実施の形態1の図4と同一のものに関しては同じ番号を付して、その説明を省略する。

【0136】

ベクトル記憶部1001は、ベクトル生成部403により生成された文書特徴ベクトルを記憶する記憶部である。この際、ベクトル生成部403において同時に生成される文書-単語行列データに付随する付加的な情報、たとえば、文書-単語行列データの行成分である単語IDとその単語との対応関係を記述した単語-単語ID対応マップデータや各単語において単語IDとその単語が有する品詞情報との対応関係を記述した単語ID-品詞対応マップデータや構文情報データなども記憶することができる。

【0137】

また、変換関数記憶部1002は、変換関数生成部404より生成された表現空間変換関数を記憶する記憶部である。

【0138】

ベクトル記憶部1001、変換関数記憶部1002は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0139】

文書特徴ベクトルと表現空間変換関数を記憶することにより記憶された文書表現空間をもちいて記憶された文書特徴ベクトルを変換することが可能になるため、ベクトル記憶部1001および変換関数記憶部1002と、ベクトル変換部405を一連の処理としておこなう必要がなくなり、機能的に分離することができる。

【0140】

つぎに、実施の形態2による文書分類装置の一連の処理の手順について説明する。図11は実施の形態2による文書分類装置の一連の処理の手順を示すフローチャートである。図11のフローチャートにおいて、実施の形態1の図8の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【0141】

ステップS830の処理につづいて、同ステップにおいて生成された文書特徴ベクトルを記憶する（ステップS831）。その後、ステップS840へ移行し、実施の形態1と同様の処理をおこなう。また、ステップS840の処理につづいて、同ステップにおいて算出された表現空間変換関数を記憶する（ステップS841）。以下、実施の形態1と同様の処理をおこなう。

【0142】

以上説明したように、実施の形態2による文書分類装置は、分類数や分類手法を変えて分類をおこなう場合に、その都度、表現空間変換関数を算出することなしに文書分類を実行できるため、複数の分類結果を短時間で得ることができる。

【0143】

さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできる。

【0144】

（実施の形態3）

さて、実施の形態1、2に対して、以下に説明する実施の形態3のように、さらにベクトル修正部1201を追加してもよい。

【0145】

まず、実施の形態3による文書分類装置の機能的構成について説明する。図12は、実施の形態3による文書分類装置の構成を機能的に示すブロック図である。図12において、実施の形態1の図4と同一のものに関しては同じ番号を付して、その説明を省略する。

【0146】

ベクトル修正部1201は、ベクトル変換部405により文書特徴ベクトルを変更する前に、解析部402により抽出される単語が有する特性により構成される規則をもちいて文書特徴ベクトル、文書特徴ベクトルを構成する特徴次元のいずれか一つまたはその両方を操作することにより文書特徴ベクトルを修正するものである。

【0147】

図13はベクトル修正部1201の処理内容の手順を示すフローチャートである。図13のフローチャートにおいて、ベクトル修正部1201は、まず、文書特徴ベクトルの読み込みをおこない（ステップS1301）、つぎに、解析部402において抽出された単語やその単語の品詞情報などを指定することで（ステップS1302）、削除などの操作をおこなう前記文書特徴ベクトルの特徴次元、すなわち文書群に固有に出現する単語IDを決定する（ステップS1303）。

【0148】

その後、ベクトル生成部403によって生成された文書特徴ベクトルやベクトル記憶部1001によって記憶された文書特徴ベクトルに対して、操作対象の特徴次元に対し、削除や合成等の修正の操作をおこない（ステップS1304）、文書特徴ベクトルを合成する。

【0149】

ベクトル修正部1201は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0150】

文書特徴ベクトルから t' 個の特徴次元（すなわち、単語 ID）を削除する手続の一例を図14に示す。文書数 d 、単語数 t とし、文書特徴ベクトル（文書 - 単語頻度行列）を $t \times d$ の大きさの行列 X とし、各行（列）が単語 ID に対応する大きさ $t \times t$ の単位行列に、削除対象となる単語 ID に対応する行を削除した $t' \times t$ の大きさの行列を P_t とした場合、修正部1201によって修正される文書特徴ベクトル X' は式10のようになる。

【0151】

$$X' = P_t X \quad (\text{式}10)$$

【0152】

つぎに、実施の形態3による文書分類装置の一連の処理の手順について説明する。図15は実施の形態1による文書分類装置の一連の処理の手順を示すフローチャートである。図13のフローチャートにおいて、実施の形態1の図8の各ステップと同じ処理をおこなうステップは同じ番号を付して、その説明を省略する。

【0153】

ステップS830、S831の処理につづいて、文書特徴ベクトルの修正をおこなう（ステップS832）。その後、ステップS840へ移行し、実施の形態1と同様の処理をおこなう。

【0154】

以上説明したように、実施の形態3による文書分類装置は、ベクトル修正部1201により、ベクトル生成部403によって文書特徴ベクトルを生成した後でも、文類時に不要であることが判明した単語などを削除することができる。さらに、同じ文書特徴ベクトルに対しての分類を効率的におこなるようになっているが、前記文書特徴ベクトル修正部1201により、各分類ごとに異なる単語を削除して文書分類を実行することができる。

【0155】

(実施の形態4)

さて、実施の形態3では、ベクトル修正部1201を追加したが、以下に説明

する実施の形態4のように、さらに、ベクトル修正部とともに変換関数修正部1601も併せて追加してもよい。

【0156】

まず、実施の形態4による文書分類装置の機能的構成について説明する。図16は、実施の形態4による文書分類装置の構成を機能的に示すブロック図である。図16において、実施の形態3の図12と同一のものに関しては同じ番号を付して、その説明を省略する。

【0157】

実施の形態3において、ベクトル修正部1201によって文書特徴ベクトルの修正がおこなわれた場合に、表現空間変換関数は修正前の文書特徴ベクトルに基づいて算出されているため、この表現空間変換関数にも文書特徴ベクトルが修正された効果を反映させなければ、文書特徴ベクトルを修正した効果が半減する可能性がある。したがって、前記表現空間変換関数を修正された文書特徴ベクトルをもとに修正する。

【0158】

すなわち、図16における変換関数修正部1601は表現空間変換関数をW'に修正する。表現空間変換関数が文書特徴ベクトルの内積に基づいて算出される場合に、表現空間変換関数は式7で与えられる。このとき、修正された表現空間変換関数をW' とすると、式2、式7、式10をもちいて式11のように表現される。

【0159】

$$W' = L^{-1} U^T P_t X (P_t X) \quad (\text{式11})$$

【0160】

変換関数修正部1601は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0161】

図17に実施の形態4による文書分類装置の一連の処理の手順を説明するフロ

ーチャートを示す。図17のフローチャートにおいて、文書特徴ベクトルの修正があった場合、表現空間変換関数の修正もおこなう（ステップS841）。以下は実施の形態3の処理と同様である。

【0162】

以上説明したように、実施の形態4による文書分類装置においては、文書特徴ベクトルの修正にともなって表現空間変換関数の修正もおこなうことができるので、より適正な文書特徴ベクトルの変換ができる。

【0163】

(実施の形態5)

さて、実施の形態4では、変換関数修正部1601を追加したが、以下に説明する実施の形態5のように、さらに変換関数修正部1601へ修正指示をおこなう変換関数修正指示部1801を追加してもよい。

【0164】

まず、実施の形態5による文書分類装置の機能的構成について説明する。図18は、実施の形態5による文書分類装置の構成を機能的に示すブロック図である。図18において、実施の形態1の図4と同一のものに関しては同じ番号を付して、その説明を省略する。

【0165】

変換関数修正指示部1801は、表現空間変換関数の特徴次元の操作に関する指示するものである。また、変換関数修正部1802は、変換関数修正指示部1801からの指示内容に基づいて、前記表現空間変換関数の特徴次元を操作し、前記表現空間変換関数を修正する。

【0166】

変換関数修正指示部1801、変換関数修正部1802は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0167】

変換関数修正指示部1801においては、操作者の意図を反映するような文書分類をおこなうための一つの方法として、前記表現空間変換関数により構成される空間における不必要的特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。

【0168】

しかしながら、表現空間変換関数により生成される空間の特徴次元は、解析部402によって抽出された単語のうち意味的に似たものが複数結合したものと考えることができるので、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものである。したがって、操作者に各特徴次元の意味を提示することは極めて難しい。

【0169】

そこで、操作者に分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することができる。

【0170】

本実施の形態では、表現空間変換関数の特徴次元を操作する例として、操作者が指定するある文書と類似度の高い特徴次元の削除をおこなう例を示す。操作者により指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに表現空間変換関数を適用し文書ベクトルを表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。

【0171】

このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などをもちいることができる。また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理もしくは判別分析などももちいることができる。

【0172】

このようにして、採用された特徴次元を表現空間変換関数から削除することで表現空間変換関数を修正することができる。この際、操作者が指示（指定）する情報としては、前記文書特徴ベクトルと同じ次元数を有するベクトル形式であれば、どのようなものでも適用可能である。

【0173】

操作者が指示する情報としては、そのほかに、より操作者にとって理解しやすいものとして、分類対象文書群以外の文書を文書特徴ベクトルと同じ次元をもつベクトルに表現したものをもちいることができる。また、操作者が指示する情報としてそのほかには、文書特徴ベクトルをもちいることができる。

【0174】

また、操作者が指示する情報としてそのほかには、解析部402によって抽出されまたは操作者が手動で入力した単語や単語品詞情報をもちいることができる。また、操作者が指示する情報としてそのほかには、分類結果記憶部407によって記憶されている事前におこなわれた分類結果である分類代表値をもちいることができる。

【0175】

上記の指示情報は、それぞれ単独でもちいるほか、それらを適切に組み合わせたものをもちいるようにしてもよい。

【0176】

図19に実施の形態5による文書分類装置の一連の処理の一部の手順を説明するフローチャートを示す。図19のフローチャートにおいて、まず、変換関数の修正の指示があるのを待って（ステップS1901肯定）、つぎに、指示の内容、すなわち、操作者が指示（指定）した指示情報をインプットする（ステップS1902）。複数の指示がある場合は、すべての指示が終了するまで同様のステップを繰り返し、指示が終了した場合（ステップS1903否定）に、インプットされた指示情報に基づいて変換関数の修正を実行し（ステップS1904）、すべての処理を終了する。

【0177】

以上説明したように、この実施の形態5によれば、表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことができる。

【0178】

(実施の形態6)

さて、実施の形態1～5に対して、以下に説明する実施の形態6のように、初期重心指定部2001および初期重心登録部2002をさらに追加するようにしてもよい。

【0179】

まず、実施の形態6による文書分類装置の機能的構成について説明する。図20は、実施の形態6による文書分類装置の構成を機能的に示すブロック図である。図20において、実施の形態1の図4と同一のものに関しては同じ番号を付して、その説明を省略する。

【0180】

初期重心指定部2001は、初期クラスタ重心を指定する指定部である。初期重心登録部2002は、初期重心指定部2001により指定された初期クラスタ重心を登録するする登録部である。また、分類部405は、初期重心登録部2002により登録された初期クラスタ重心にしたがって文書を分類するものである。

【0181】

初期重心指定部2001、初期重心登録部2002は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0182】

通常、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等をもちいて文書分類をおこなう場合にもちいられる分類基準が統計的な理論を元にして構成されている。しかしながら、本実施の形態においては、文書分類をおこな

った際の最終的な分類の質の評価は、統計的な数値評価ではなく、その分類結果を分析する操作者による主観評価となる。したがって、前記文書分類をおこなうための諸手法において、操作者が介入できうる余地を設けることで分類結果に操作者の意図を反映することができ、結果として分類結果の質的な向上が見込まれる。

【0183】

つぎに、図21に実施の形態6による文書分類装置の一連の処理の一部の手順を説明するフローチャートを示す。非階層型のクラスタリング手法は一般的に図19のフローチャートのような処理の手順となる。図21のフローチャートにおいて、まず、初期クラスタ重心が指定され（ステップS2101）、その初期クラスタ重心が登録される（ステップS2102）。つぎに、初期クラスタ重心を決定し（ステップS2103）、そのクラスタ重心と各分類対象データとの類似度を計算する（ステップS2104）。

【0184】

つぎに、各分類対象データを一番類似度の高いクラスタに割り当てて（ステップS2105）、各クラスタごとに割り当てられた分類対象データを基にそのクラスタ重心を計算する（ステップS2106）。

【0185】

この時点で、反復停止基準を満たすか否かを判断し（ステップS2107）、反復停止基準を満たさない場合（ステップS2107否定）は、ステップS2104へ移行し、以後、ステップS2104～S2106の各ステップを繰り返し実行する。ステップS2107において、反復停止基準を満たす場合（ステップS2107肯定）は、すべての処理終了する。

【0186】

分類結果はどのような初期クラスタ重心を選択するかに強く依存するといわれている。したがって、分類実行部での分類手法として、k-means法などの非階層型クラスタリング手法をもちいて、その初期クラスタ重心を操作者が指定することで、操作者の分類手続きへの介入を可能にし、操作者の意図を反映した文書分類が実現できる。

【0187】

なお、各文書特徴ベクトルとクラスタの重心ベクトルとの類似度を算出し、各特徴ベクトルで最も類似度の高い分類代表値にその文書特徴ベクトルを帰属させる形式の分類手法であれば、非階層型クラスタリング以外の手法でも利用可能である。また、クラスタの重心ベクトルと文書ベクトルとの類似度を測るための類似測度としては、余弦測度、内積測度、ユークリッド距離測度、マハラノビス距離測度などが利用可能である。

【0188】

初期重心指定部2001によって、前記分類対象データと同一の特徴次元数をもつ任意の複数の文書ベクトルがクラスタリングの初期重心として入力される。前記任意の文書ベクトルは操作者により指定することもできるし、また分類対象の文書特徴ベクトルなどに基づいて構築した規則を操作者が選択することにより間接的に文書ベクトルを指定することもできる。

【0189】

また、前記任意の文書ベクトルとしては、前記文書特徴ベクトルと同じ次元数を有するベクトル形式であれば、どのようなものでも適用可能である。また、任意の文書ベクトルとしては、そのほかに、より操作者にとって理解しやすいものとして、分類対象文書群以外の文書を文書特徴ベクトルと同じ次元をもつベクトルに表現したものをもちいることができる。

【0190】

また、任意の文書ベクトルとしてそのほかには、文書特徴ベクトルをもちいることができる。任意の文書ベクトルとしてそのほかには、解析部402によって抽出される単語や単語品詞情報をもちいることができる。また、任意の文書ベクトルとしてそのほかには、分類結果記憶部407によって記憶されている事前におこなわれた分類結果である分類代表値をもちいることができる。

【0191】

上記の指示情報は、それぞれ単独でもちいるほか、それらを適切に組み合わせたものをもちいるようにしてもよい。

【0192】

2つの任意の文書ベクトル、3つの文書特徴ベクトル、一つの単語、一つの分類代表値とそれらの組み合わせ規則を指定することで、5つの初期クラスタ重心を求める例を図22に示す。図22に示すとおり、本実施の形態では、初期クラスタ重心1として文書1を、初期クラスタ重心2として文書2と文書3の平均を、初期クラスタ重心3として文書4と単語1の平均を、初期クラスタ重心4として文書5を、初期クラスタ重心5として分類代表値1を各々指定している。

【0193】

また、指定された文書ベクトルが、操作者が指定したクラスタ数に満たない場合には、*k-means*法などでもちいれられている一般的な自動初期重心選出法をもちいて残りのクラスタ重心を求めることができる。このようにして求めた初期重心に基づいて*k-means*法等をもちいて、クラスタの精練化をおこなうことで文書分類を実行する。

【0194】

以上説明したように、この実施の形態6によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことができる。

【0195】

【発明の効果】

以上説明したように、請求項1の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能な文書分類装置が得られるという効果を奏する。

【0196】

また、請求項2の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類装置が得

られるという効果を奏する。

【0197】

また、請求項3の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0198】

また、請求項4の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0199】

また、請求項5の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0200】

また、請求項6の発明によれば、表現空間変換関数が文書特徴ベクトルの内積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるるので、より適正な文書特徴ベクトルの変換をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0201】

また、請求項7の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0202】

また、請求項8の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0203】

また、請求項9の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0204】

また、請求項10の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0205】

また、請求項11の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0206】

また、請求項12の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に

指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0207】

また、請求項13の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0208】

また、請求項14の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0209】

また、請求項15の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【0210】

また、請求項16の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0211】

また、請求項17の発明によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することが可能な文書分類

方法が得られるという効果を奏する。

【0212】

また、請求項18の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積をもちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0213】

また、請求項19の発明によれば、表現空間変換関数を導出する際に必要となる文書間の類似性として文書特徴ベクトル間の内積に加え、文書の作成者や作成日などの文書間類似情報ももちいることにより、文書間の意味的な関連性を反映した文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0214】

また、請求項20の発明によれば、算出する文書特徴ベクトルと表現空間変換関数を記憶することにより、表現空間変換関数を算出する部分と実際に前記表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分とを分離して処理するので、その都度、表現空間変換関数を算出することなしに文書分類を実行でき、さらに、前記文書特徴ベクトル変換部でもちいる表現空間変換関数として、事前に他の文書特徴ベクトルに基づいて生成された表現空間変換関数をもちいることもできるため、文書分類の繰り返し実行を短時間で効率良くおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0215】

また、請求項21の発明によれば、文書分類の繰り返し実行をおこなう際、個々の分類実行ごとに、文書特徴ベクトルやそれらを構成する特徴次元を操作することで、各分類ごとに異なる単語を削除して文書分類を実行する等の分類対象文書の範囲の変更や分類をおこなう空間の変更をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0216】

また、請求項22の発明によれば、表現空間変換関数が文書特徴ベクトルの内

積をに基づいて算出される場合、表現空間変換関数をもちいて変換された文書をもちいて文書分類をおこなう部分において、文書特徴ベクトルやその特徴次元が操作された場合に生じる表現空間変換関数の不整合を簡便に修正することができるので、より適正な文書特徴ベクトルの変換をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0217】

また、請求項23の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について操作者が簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0218】

また、請求項24の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された分類対象以外の任意の文書ベクトルデータをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0219】

また、請求項25の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元について、操作者により指示された文書特徴ベクトルをもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0220】

また、請求項26の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された解析情報をもちいての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0221】

また、請求項27の発明によれば、前記表現空間変換関数をもちいて構成される空間の特徴次元を、操作者により指示された事前に分類された分類結果をもち

いての簡便な操作をすることにより、操作者の意図を反映しうる文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0222】

また、請求項28の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心を、操作者が任意に指定することができ、その指定された初期クラスタ重心にしたがって文書分類をおこなうので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0223】

また、請求項29の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象以外の任意の文書をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0224】

また、請求項30の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、文書特徴ベクトルをもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0225】

また、請求項31の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、分類対象文書を文書解析部に作用させた結果得られる単語等の解析情報をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0226】

また、請求項32の発明によれば、文書分類手法として、非階層型クラスタリング手法をもちいて、その際に必要となる初期クラスタ重心として、事前に分類された分類結果をもちいることができるので、操作者の意図を反映する文書分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【0227】

また、請求項33の発明によれば、請求項17~32に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項17~32の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図1】

この発明の実施の形態1による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【図2】

実施の形態1による文書分類装置を構成する情報処理システムにおけるサーバーをハードウェア的に示す説明図である。

【図3】

実施の形態1による文書分類装置を構成する情報処理システムにおけるクライアントをハードウェア的に示す説明図である。

【図4】

実施の形態1による文書分類装置の構成を機能的に示すブロック図である。

【図5】

実施の形態1による文書分類装置の構成を機能的に示す別のブロック図である

【図6】

実施の形態1による文書分類装置の構成を機能的に示す別のブロック図である

【図7】

実施の形態1による文書分類装置の文書-単語行列データと文書特徴ベクトルの一例を示す説明図である。

【図8】

実施の形態1による文書分類装置の一連の処理の手順を示すフローチャートである。

【図9】

実施の形態1による文書分類装置の一連の処理の別の手順を示すフローチャートである。

【図10】

この発明の実施の形態2による文書分類装置の構成を機能的に示すブロック図である。

【図11】

実施の形態2による文書分類装置の一連の処理の手順を示すフローチャートである。

【図12】

この発明の実施の形態3による文書分類装置の構成を機能的に示すブロック図である。

【図13】

実施の形態3による文書分類装置のベクトル修正部の処理内容の手順を示すフローチャートである。

【図14】

実施の形態3による文書分類装置の文書特徴ベクトルから特徴次元を削除する手続きの一例を示す説明図である。

【図15】

実施の形態3による文書分類装置の一連の処理の手順を示すフローチャートである。

【図16】

この発明の実施の形態4による文書分類装置の構成を機能的に示すブロック図である。

【図17】

実施の形態4による文書分類装置の一連の処理の手順を示すフローチャートである。

【図18】

この発明の実施の形態5による文書分類装置の構成を機能的に示すブロック図

である。

【図19】

実施の形態5による文書分類装置の一連の処理の一部の手順を示すフローチャートである。

【図20】

この発明の実施の形態6による文書分類装置の構成を機能的に示すブロック図である。

【図21】

実施の形態6による文書分類装置の一連の処理の一部の手順を示すフローチャートである。

【図22】

実施の形態6による文書分類装置の初期クラスタ重心を求める処理の一例についての説明図である。

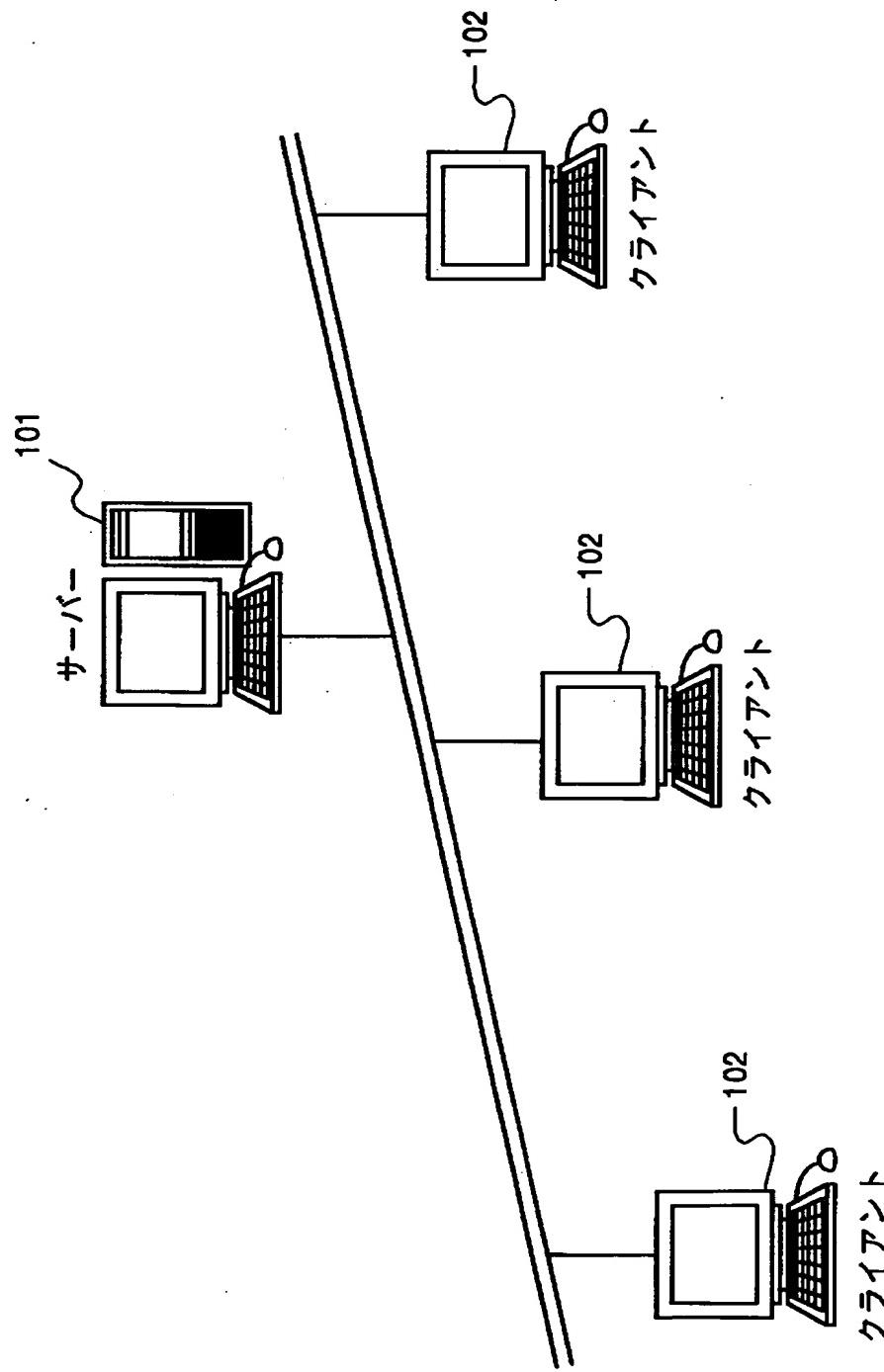
【符号の説明】

- 101 サーバー
- 102 クライアント
- 103 ネットワーク
- 201 CPU
- 204 I/F
- 206 ディスク装置
- 301 CPU
- 306 ハードディスク
- 308 ディスプレイ
- 309 I/F
- 311 キーボード
- 312 マウス
- 313 スキャナ
- 401 入力部
- 402 解析部

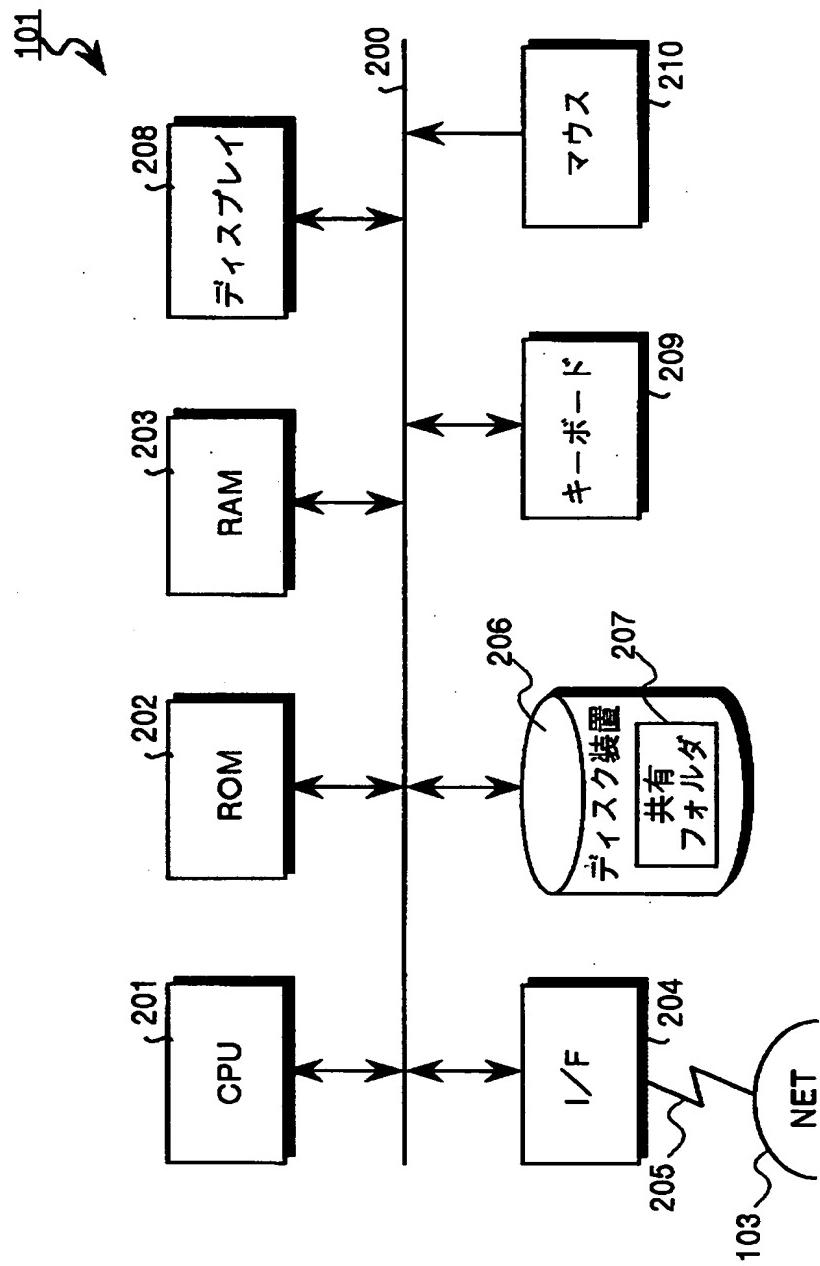
403 ベクトル生成部
404 変換関数算出部
405 ベクトル変換部
406 分類部
407 分類結果記憶部
421 内積算出部
431 文書間類似情報設定部
1001 ベクトル記憶部
1002 変換関数記憶部
1201 ベクトル修正部
1601 変換関数修正部
1801 変換関数修正指示部
2001 初期重心指定部
2002 初期重心登録部

【書類名】 図面

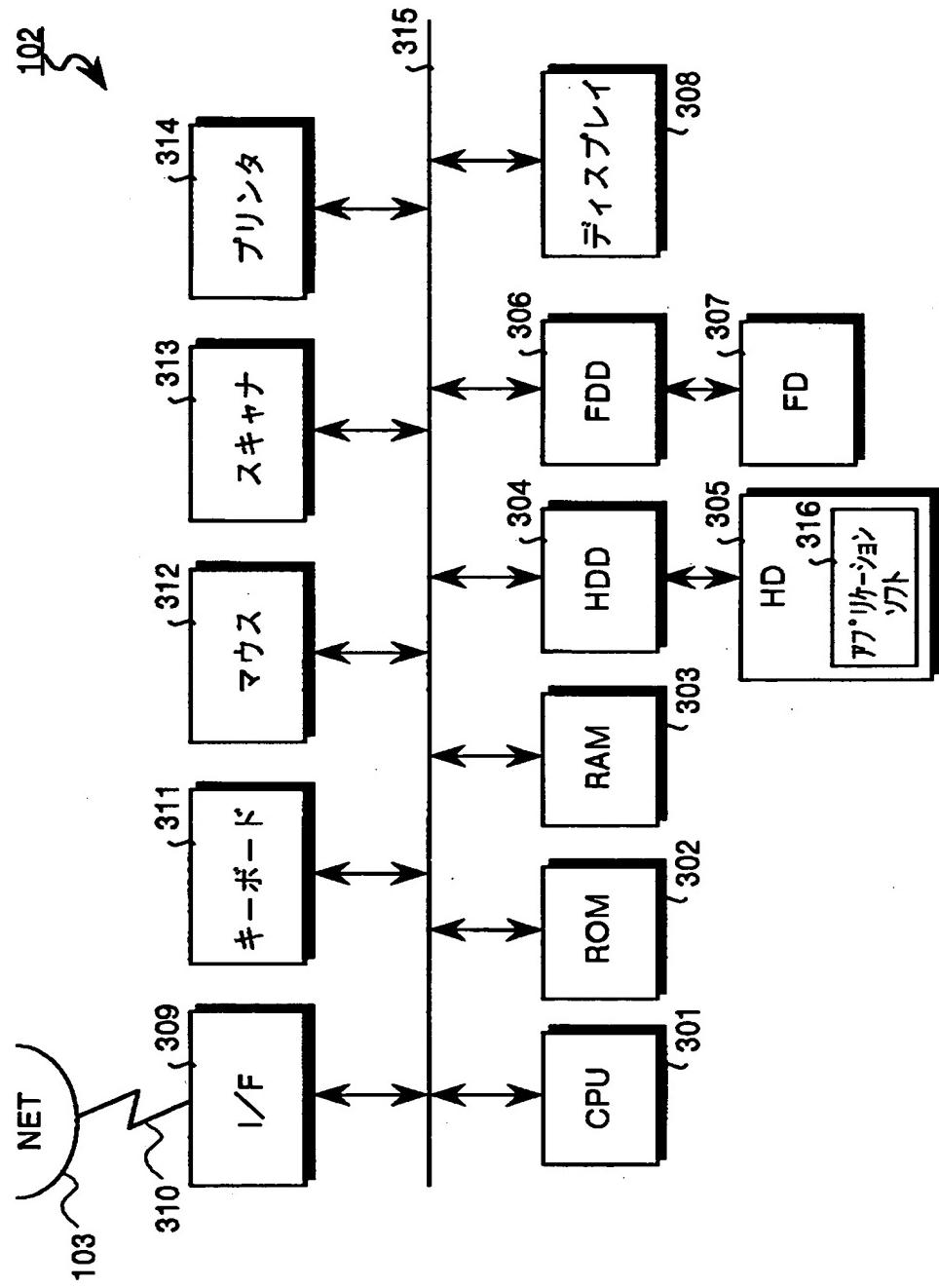
【図1】



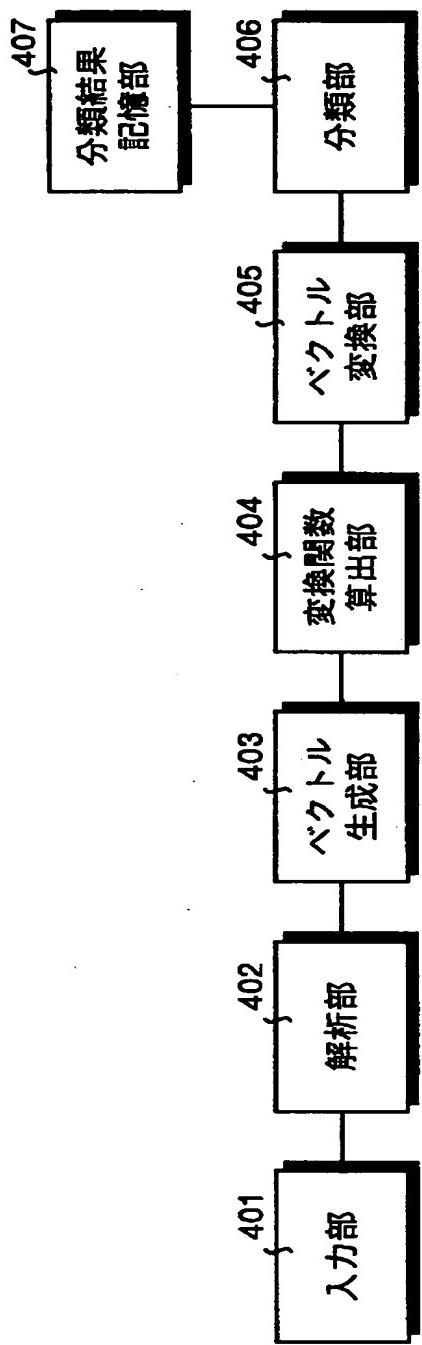
【図2】



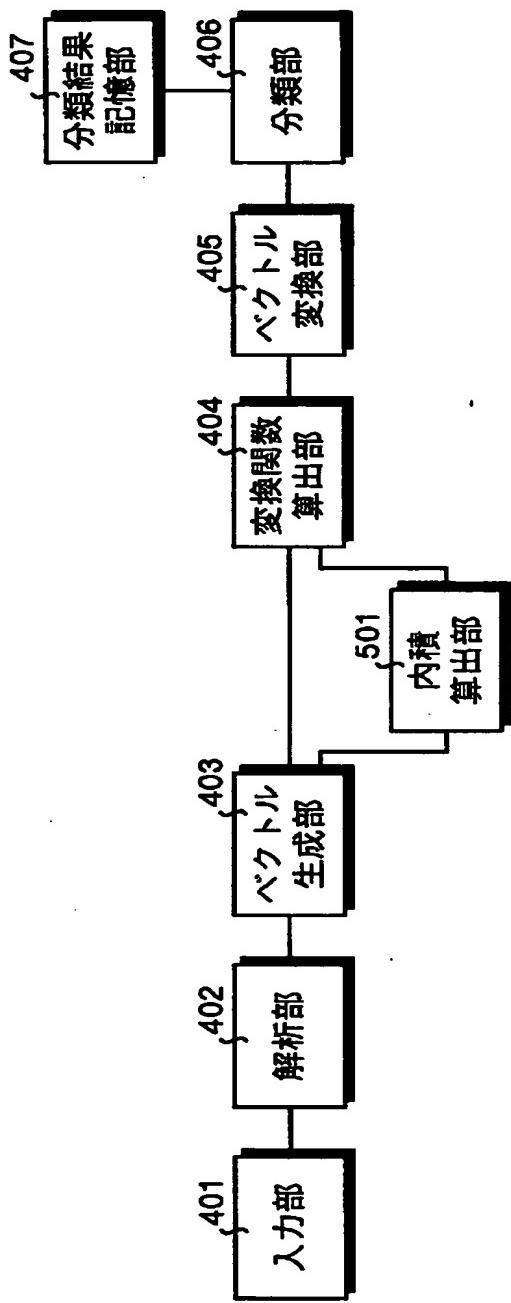
【図3】



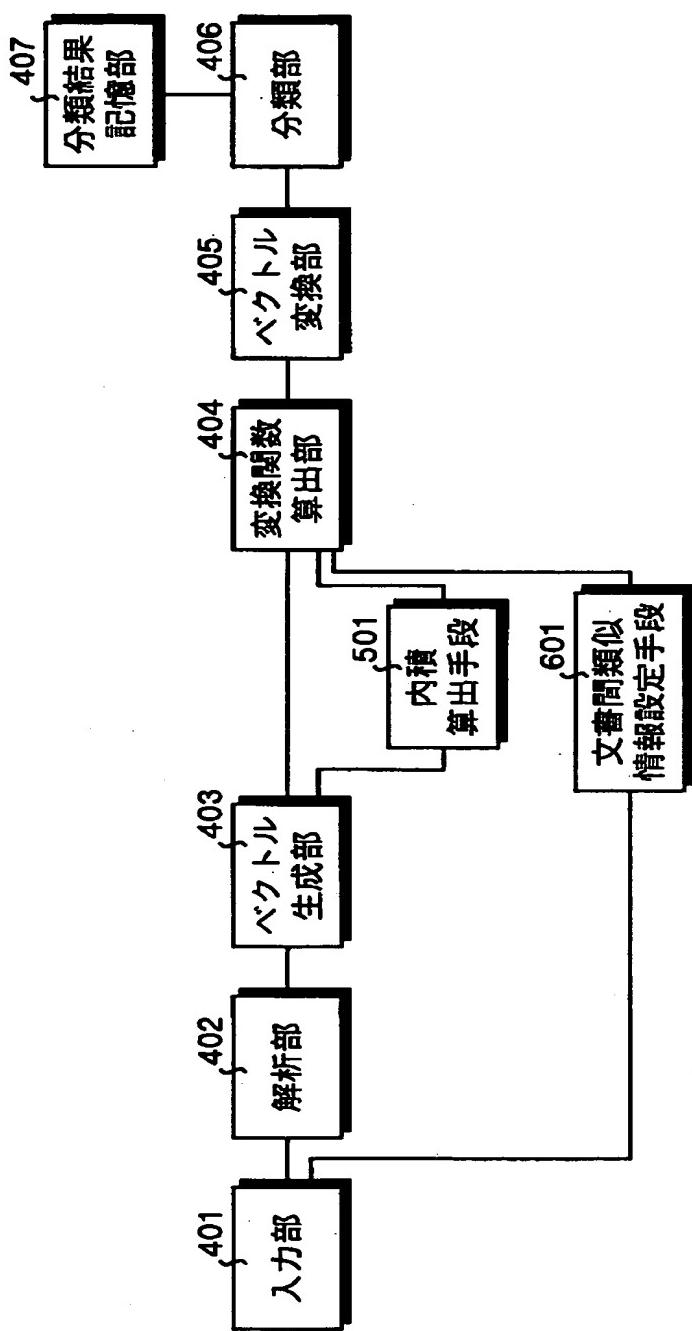
【図4】



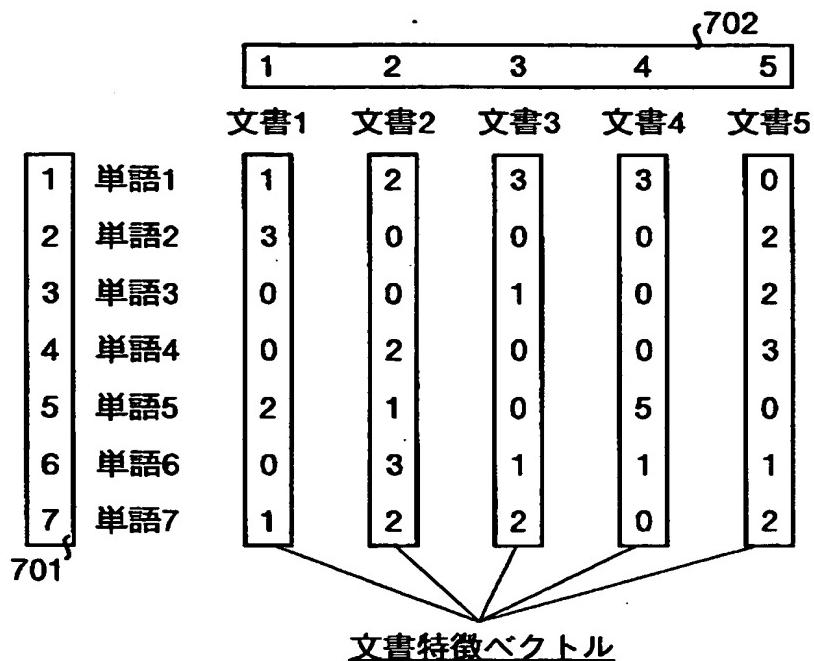
【図5】



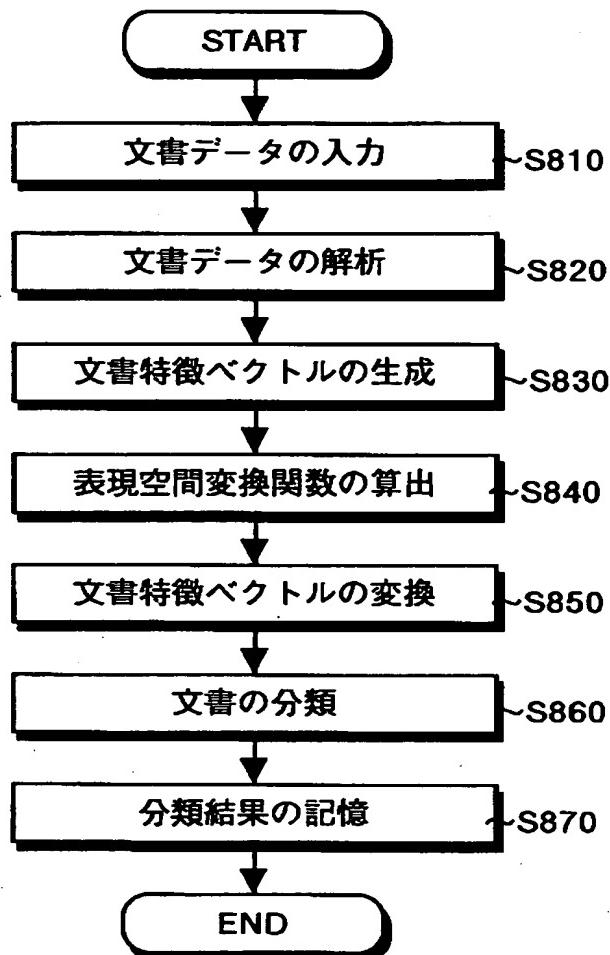
[図6]



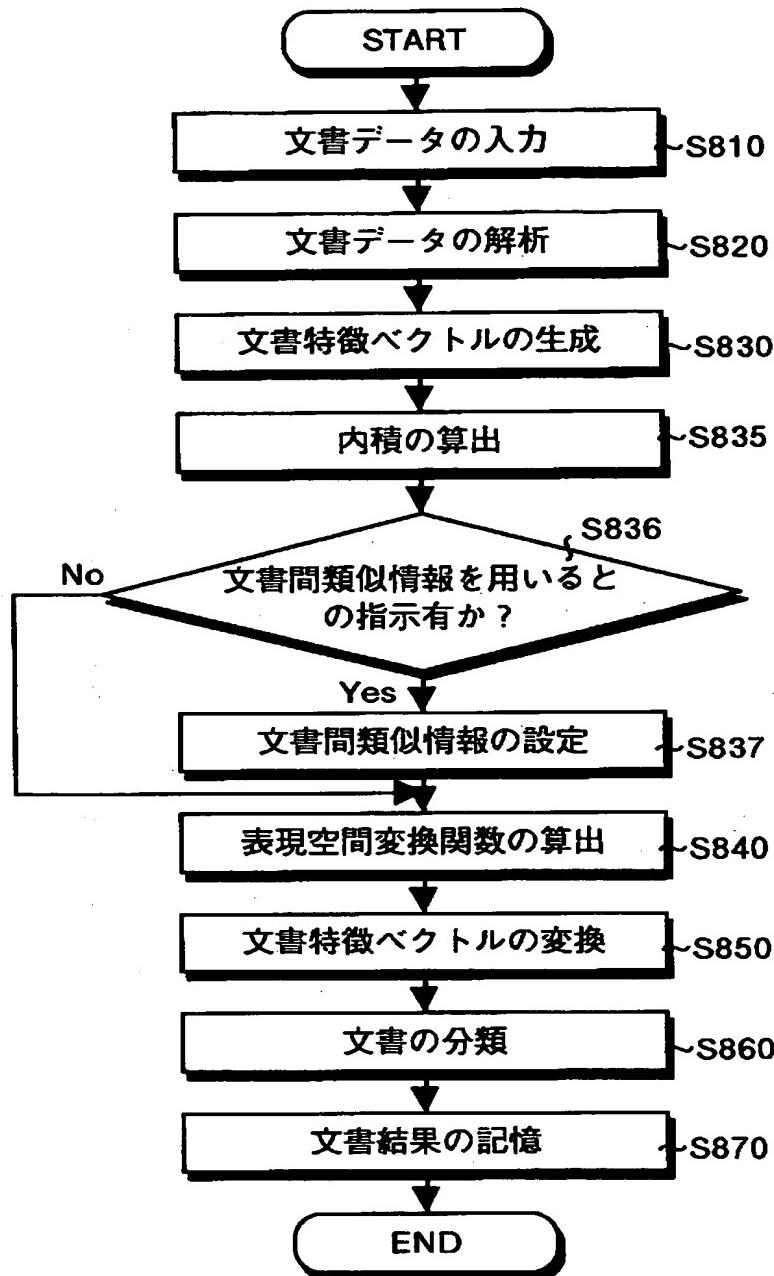
【図7】



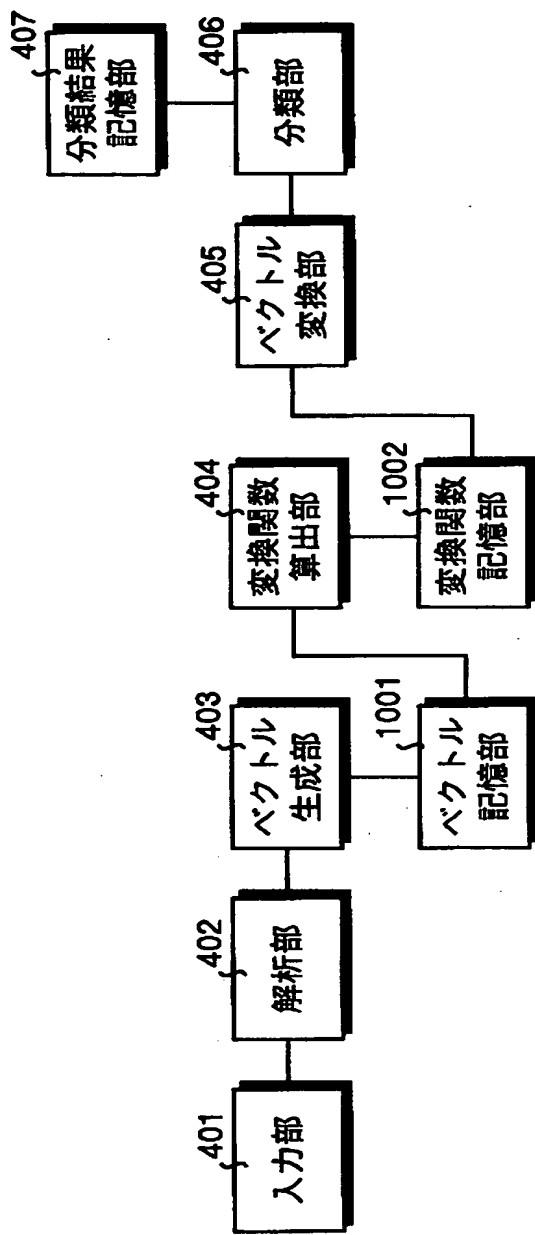
【図8】



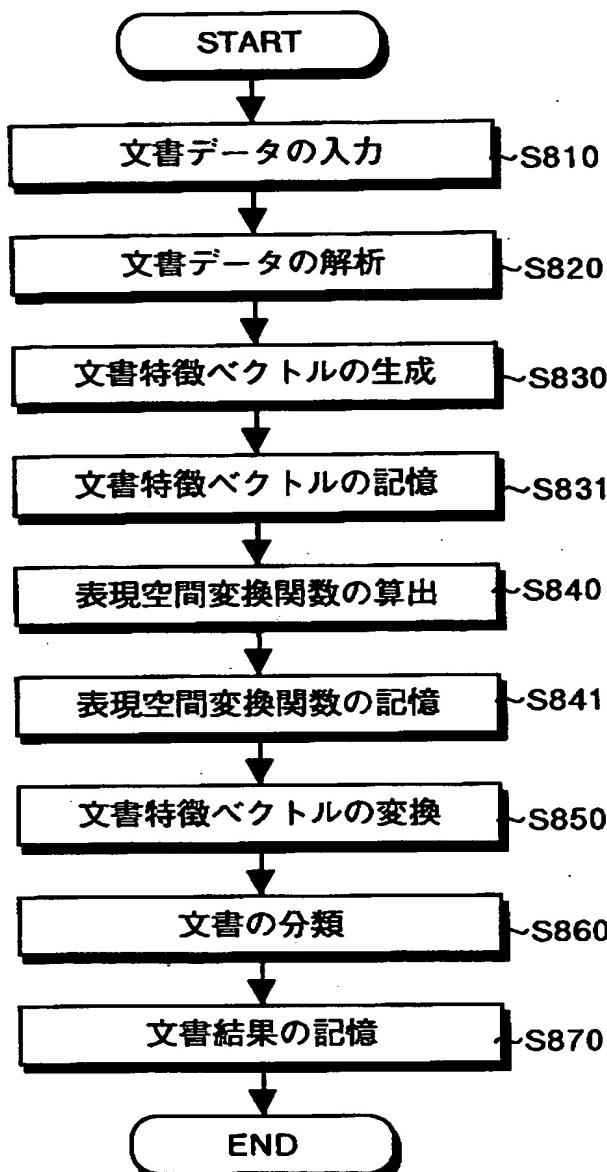
【図9】



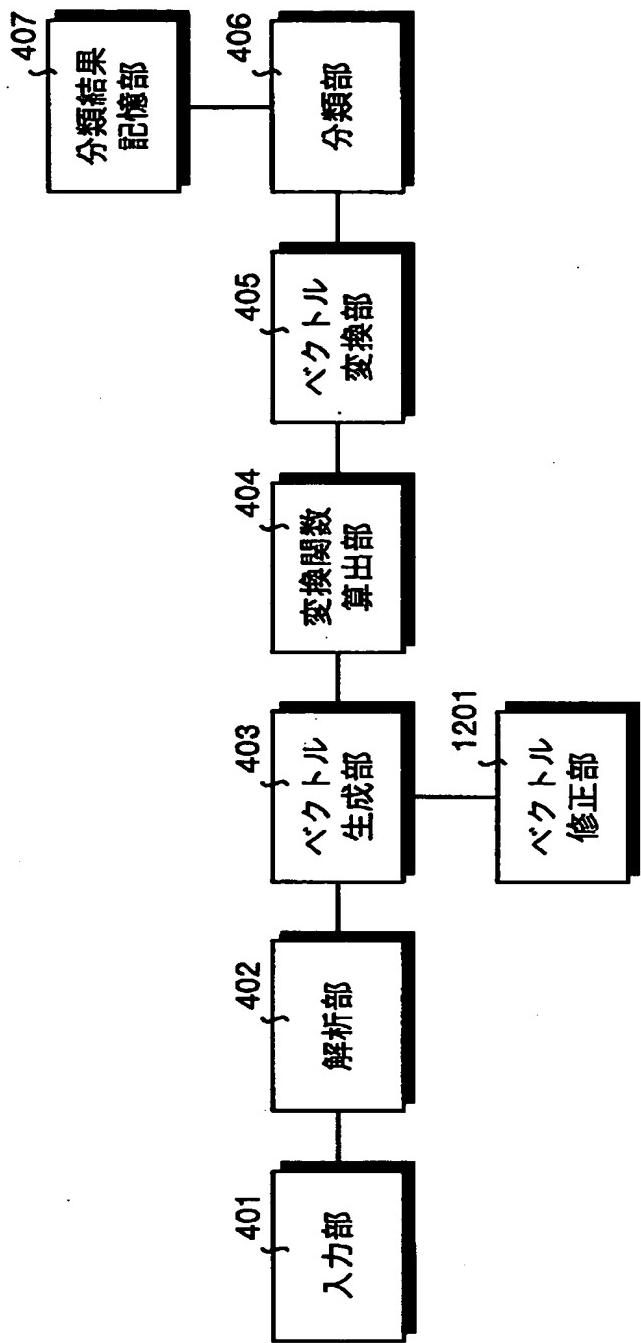
【図10】



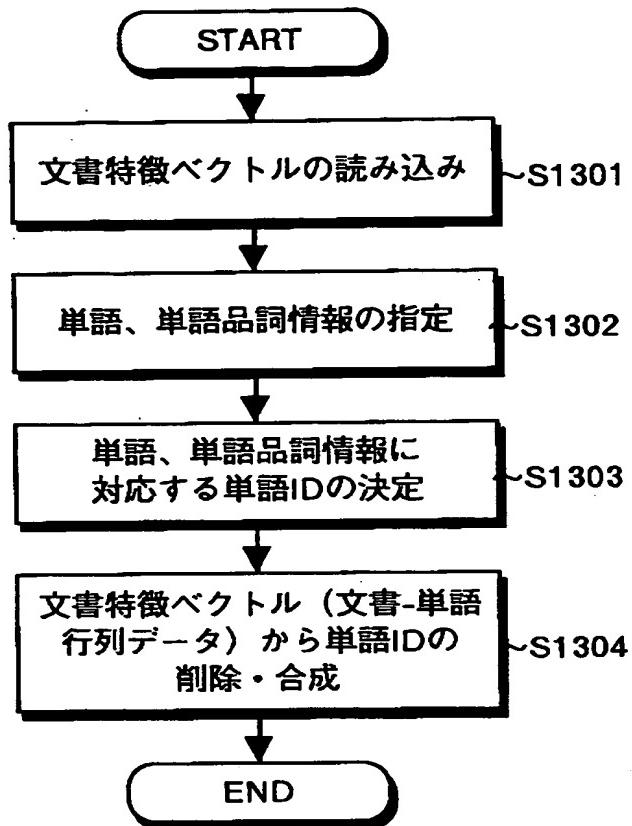
【図11】



【図12】



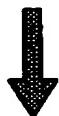
【図13】



【図14】

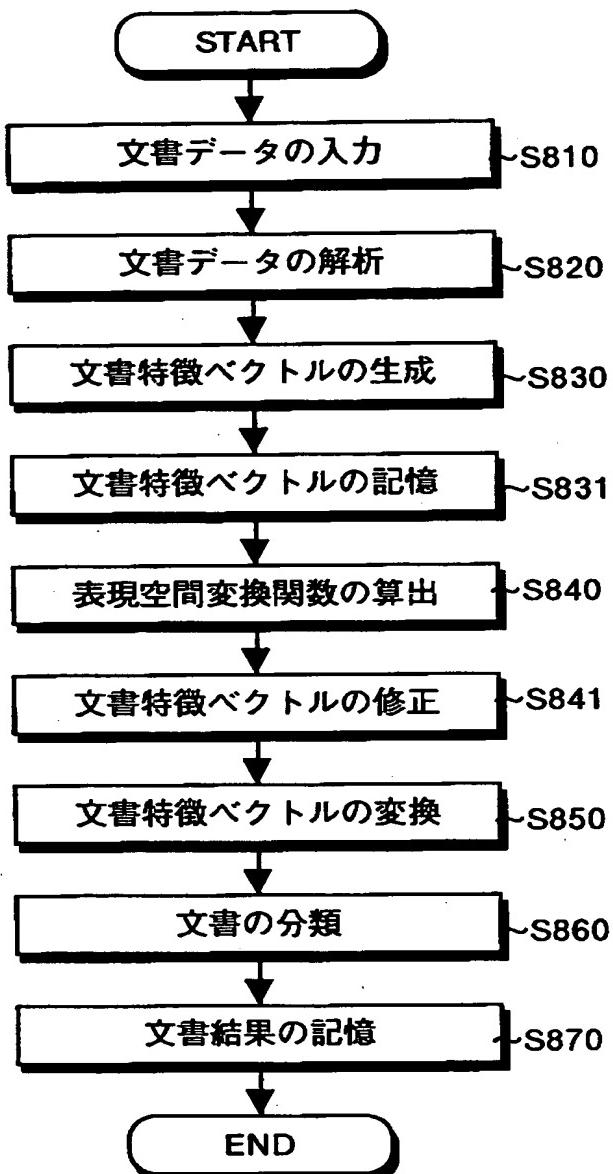
単語2と単語5を削除

	文書1	文書2	文書3	文書4	文書5
1 単語1	1	2	3	3	0
2 単語2	3	0	0	0	2
3 単語3	0	0	1	0	2
4 単語4	0	2	0	0	3
5 単語5	2	1	0	5	0
6 単語6	0	3	1	1	1
7 単語7	1	2	2	0	2

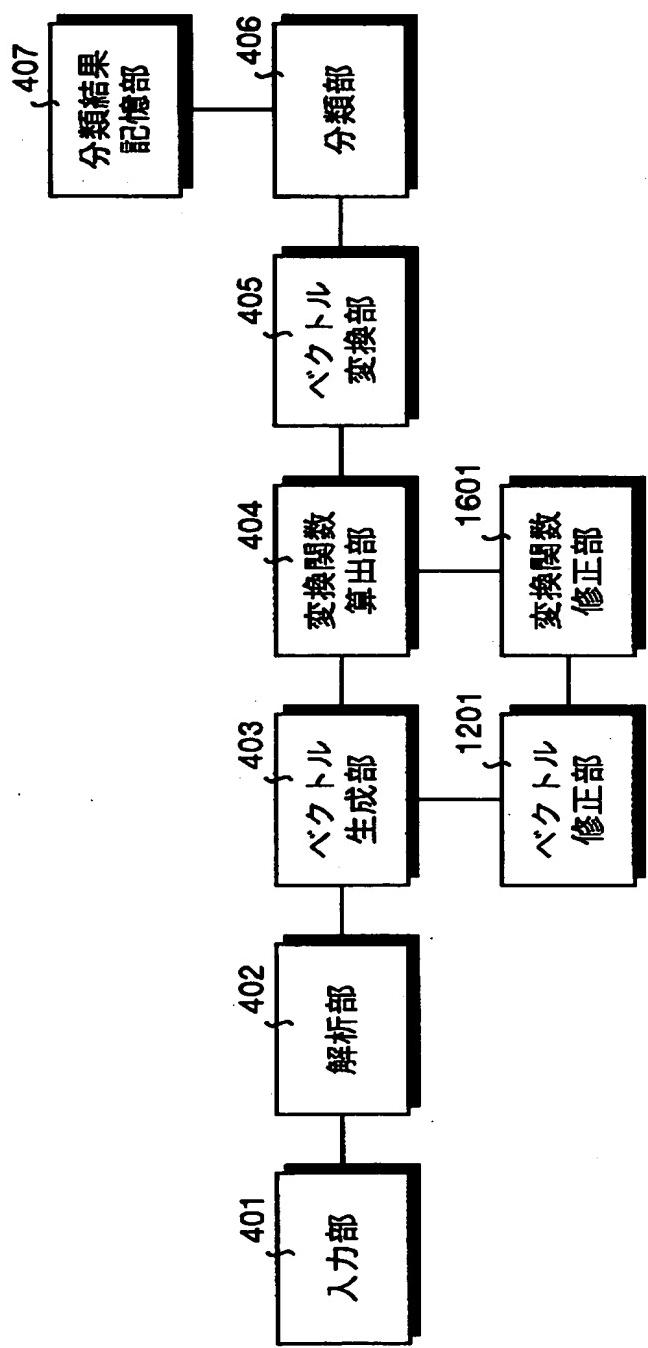


	文書1	文書2	文書3	文書4	文書5
1 単語1	1	2	3	3	0
3 単語3	0	0	1	0	2
4 単語4	0	2	0	0	3
6 単語6	0	3	1	1	1
7 単語7	1	2	2	0	2

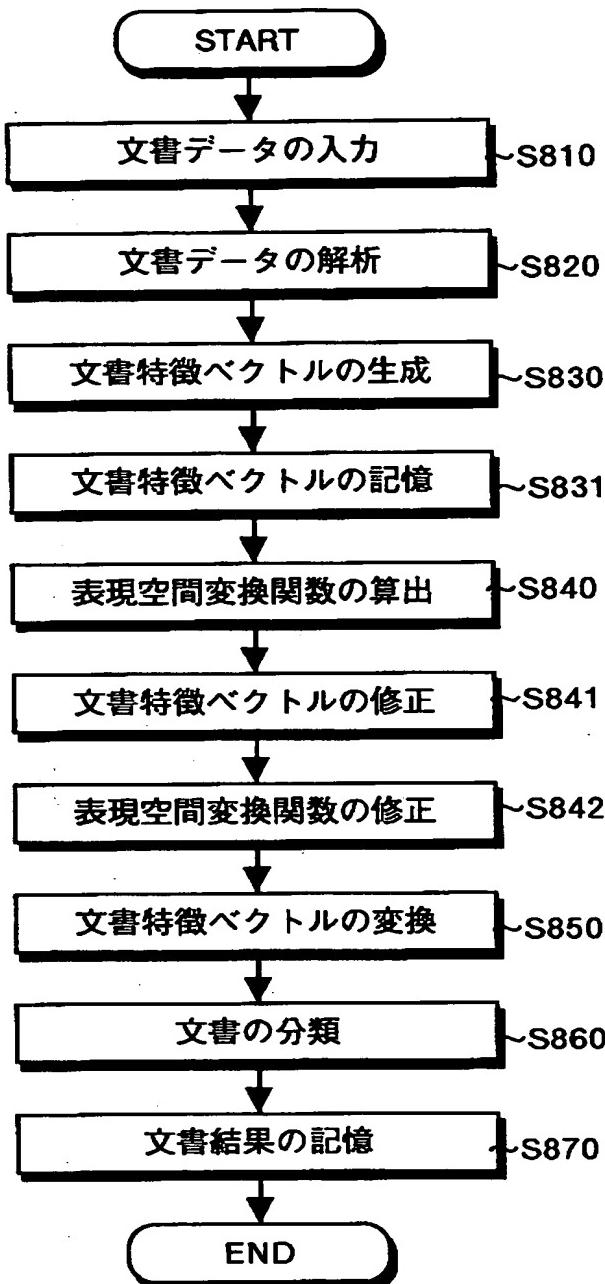
【図15】



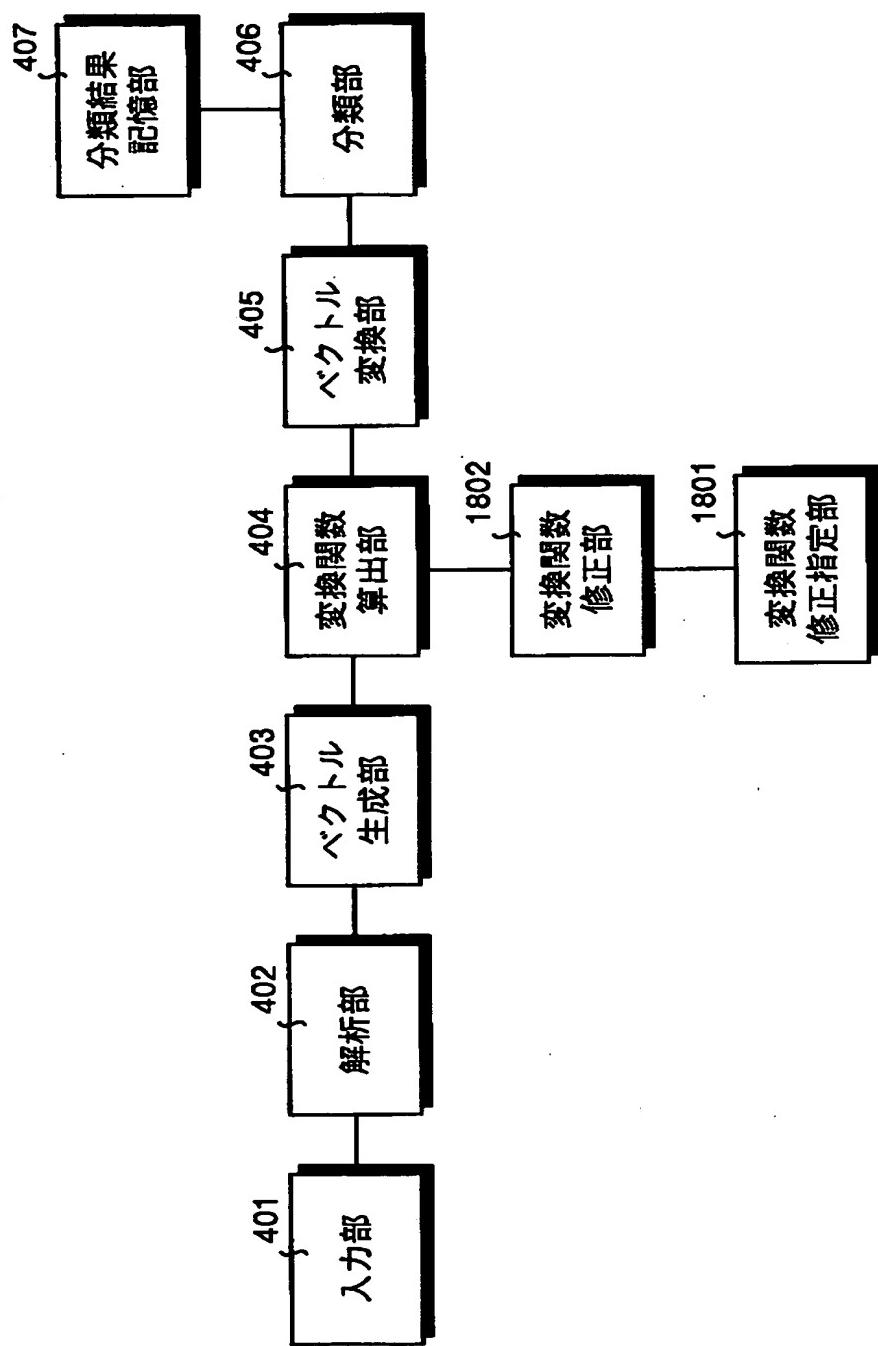
【図16】



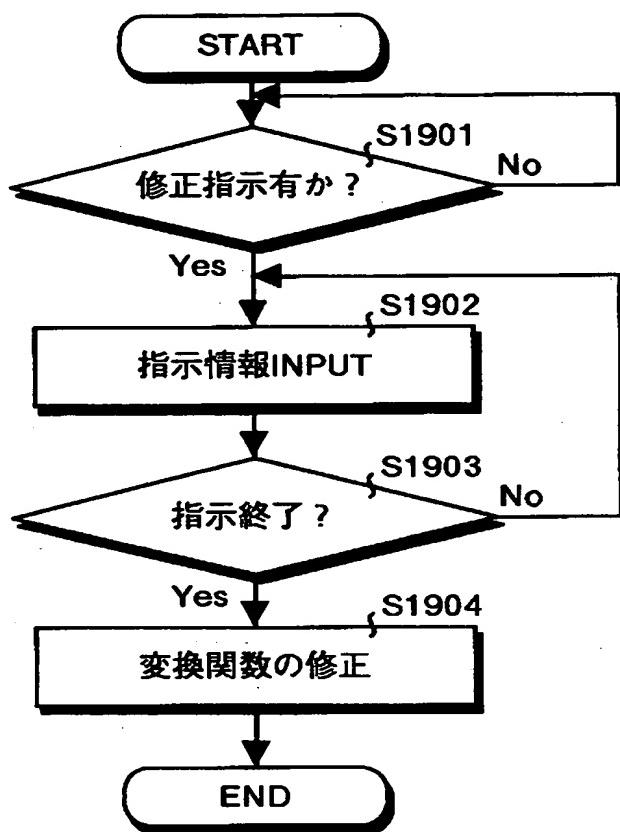
【図17】



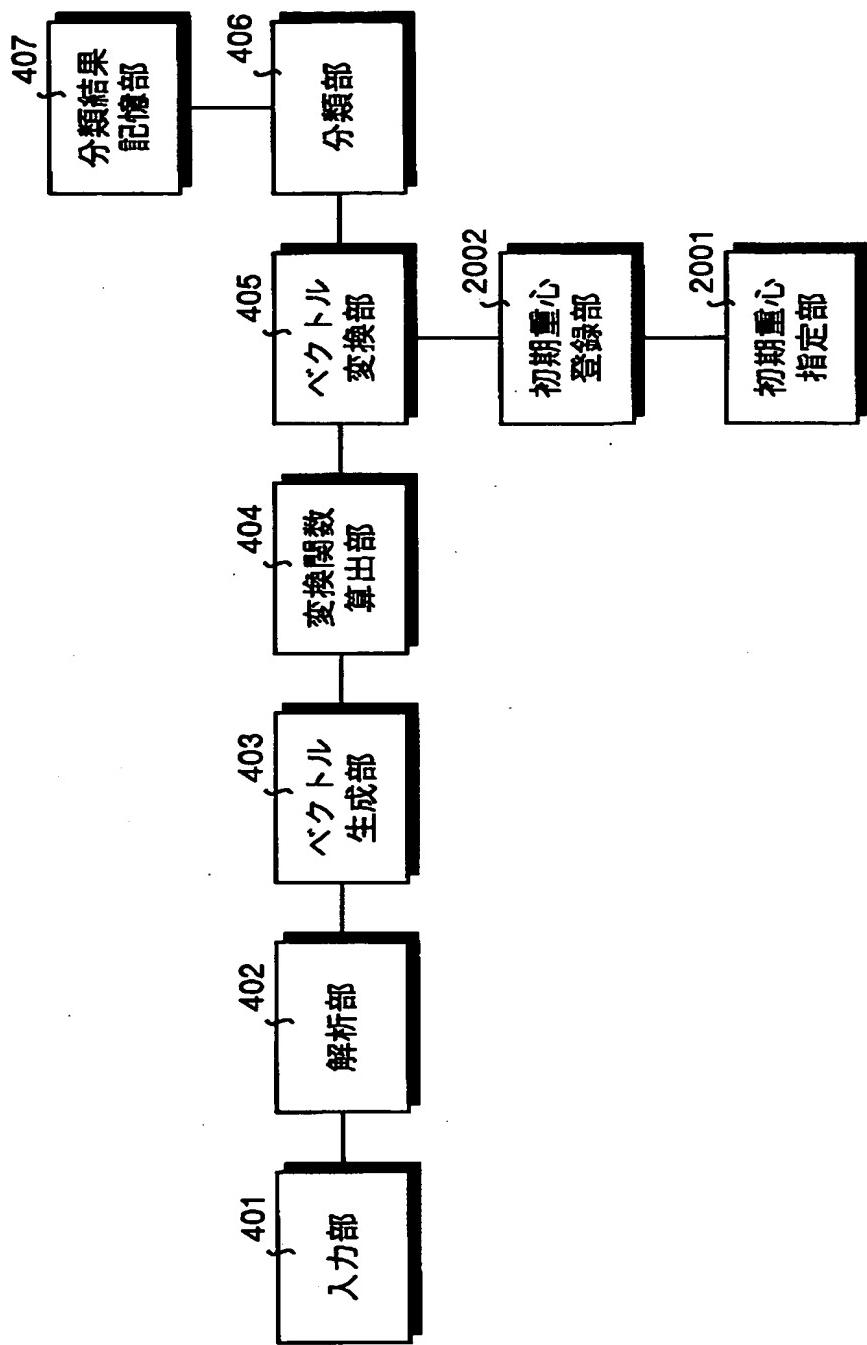
[図18]



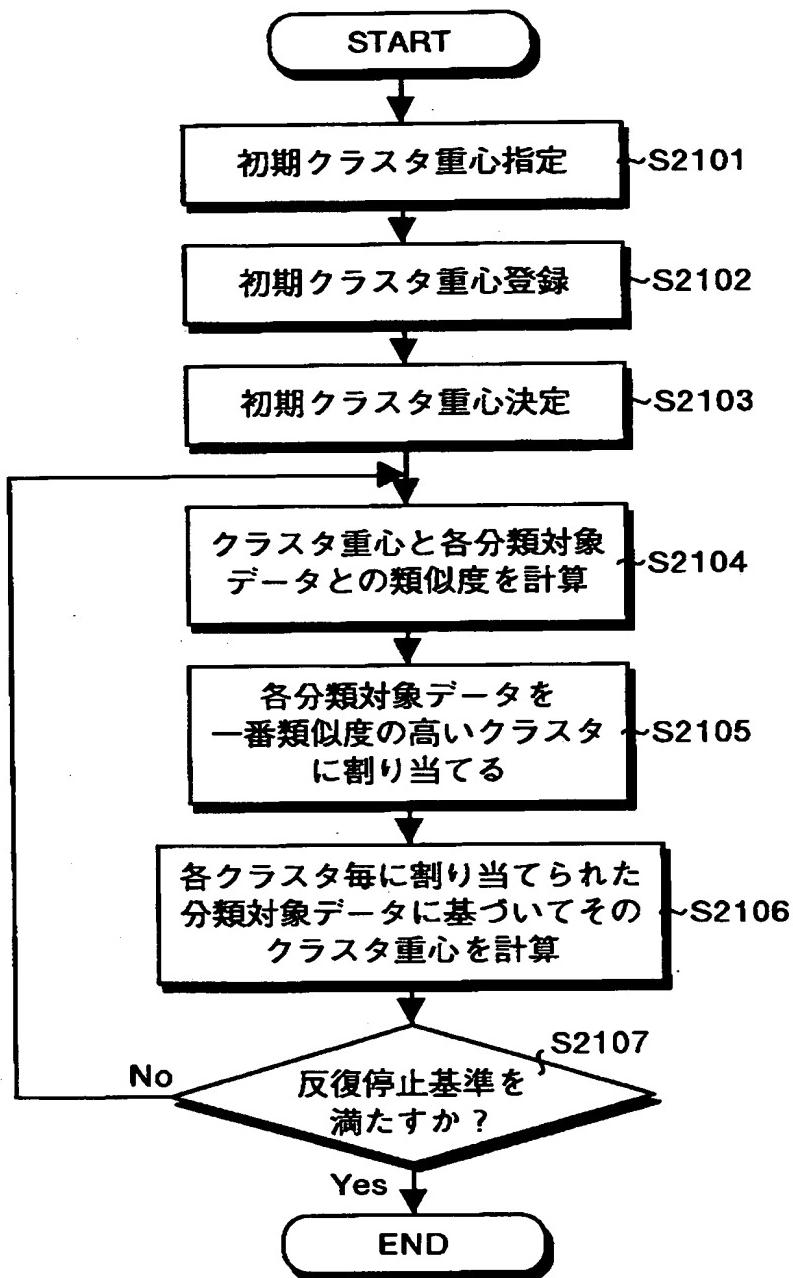
【図19】



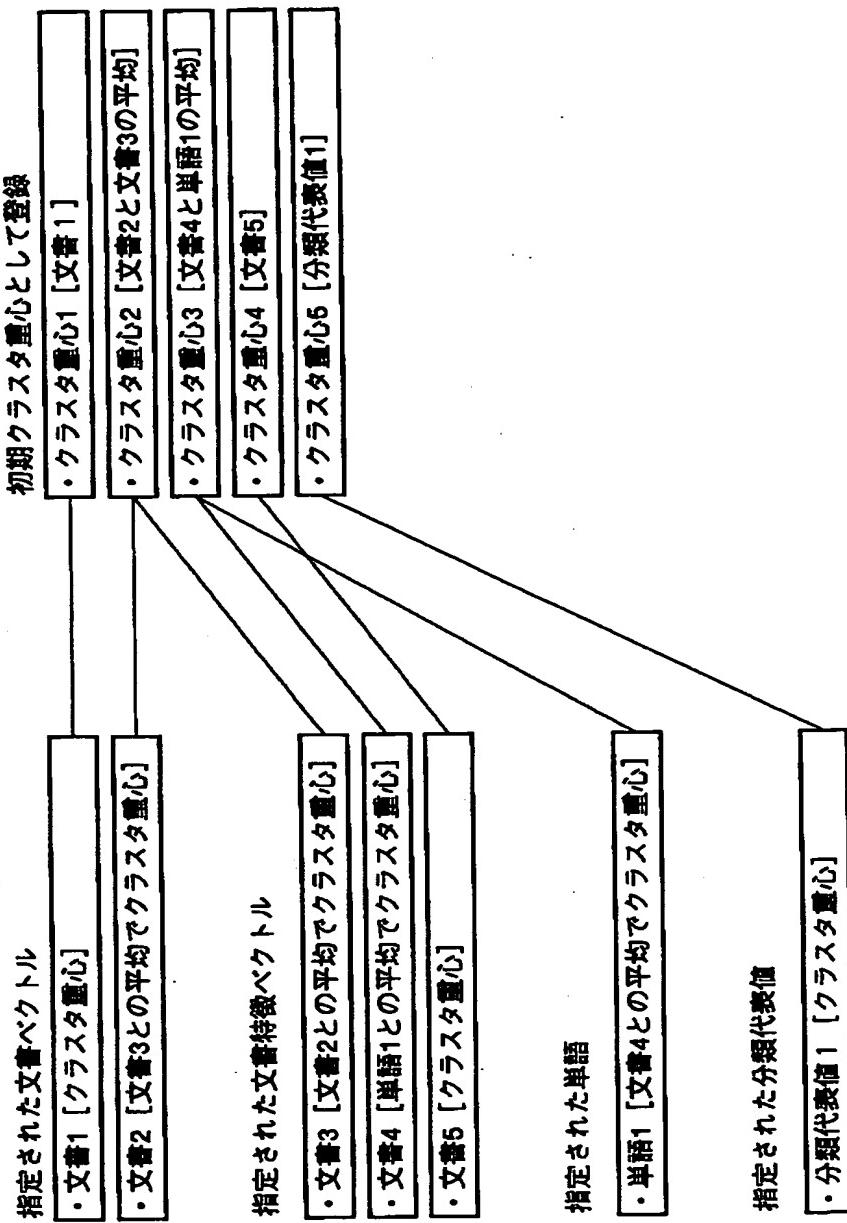
【図20】



【図21】



【図22】



【書類名】 要約書

【要約】

【課題】 文書間の類似性に基づいて文書分類をおこなう際、操作者の意図を反映する文書分類を短時間で効率良く繰り返しをおこなうことを課題とする。

【解決手段】 文書データを入力する入力部401と、入力された文書データを解析し解析情報を得る解析部402と、得られた解析情報に基づいて文書データに対する文書特徴ベクトルを生成するベクトル生成部403と、生成された文書特徴ベクトルが文書特徴ベクトル間の類似性を反映する空間に射影されるための表現空間変換関数を算出する変換関数算出部404と、算出された表現空間変換関数をもちいてベクトル生成部403により生成された文書特徴ベクトルを変換するベクトル変換部405と、変換された文書特徴ベクトル間の類似度に基づいて文書を分類する分類部406と、分類された文書分類の結果を記憶する分類結果記憶部407とを備える。

【選択図】 図4

【書類名】 職権訂正データ
【訂正書類】 特許願

<認定情報・付加情報>

【特許出願人】 申請人
【識別番号】 000006747
【住所又は居所】 東京都大田区中馬込1丁目3番6号
【氏名又は名称】 株式会社リコー

出願人履歴情報

識別番号 [000006747]

1. 変更年月日 1990年 8月24日

[変更理由] 新規登録

住 所 東京都大田区中馬込1丁目3番6号

氏 名 株式会社リコー